GENETICS OF PLASMINOGEN ACTIVATOR INHIBITOR – 1: A POTENT BIOLOGICAL EFFECTOR OF CARDIOVASCULAR DISEASE RISK

By

Marquitta Jonisse White

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

December, 2014

Nashville, Tennessee

Approved:

Scott M. Williams, Ph.D.
Melinda C. Aldrich, Ph.D.
Dana C. Crawford, Ph.D.
Bingshan Li, Ph. D.
Jason H. Moore, Ph.D.
Nancy J. Brown, M.D.

To Mimi, for giving me more than I could ever repay

To Godfather, for always encouraging me to follow my dreams

and

To my fiancé, Garric F. D. Smith, for being my Shield

ACKNOWLEDGEMENTS

This work was made possible by funding from the Center for Human Genetics Research at Vanderbilt University and the Institute for Quantitative Biomedical Sciences at Dartmouth College, as well as grants under the direction of Scott M. Williams and Jason Moore. I would like to thank my committee chair Bingshan Li and the members of my thesis committee Nancy Brown, Jason Moore, and particularly, my two Vanderbilt co-mentors Dana Crawford and Melinda Aldrich for their support and guidance. I would especially like to recognize and thank my primary mentor, Scott Williams, for his unwavering enthusiasm and support of my development into an independent researcher. He gave me the training and exposure to new ideas and collaborations that enhanced not only my repertoire as a statistical geneticist but also prepared me to be an active member of the scientific community. I would also like to thank my fellow Williams lab members, and the professors and students in the Human Genetics program for their support. In particular, I would like to thank Nuri Kodaman, Rafal Sobota, and Jackie Bartlett for their friendship, support, guidance, and advice.

I would like to acknowledge my parents, Michael and Verice White, who have shown by example the rewards that come from perseverance and determination. I would like to particularly thank them for giving me my “village” of countless aunts, uncles, cousins, and extended family that have been my support system from day one. I would like to especially thank my three siblings for their love, strength, kindness, and unending faith in me. My sisters, Aisha and Viana, have always been there to love, support, challenge, tease, and torment me and I would be lost without them. And last, but certainly not least, I would like to acknowledge my future husband, Garric Smith. From the first day that we met, he challenged me to reach for the stars and dared me to chase my dreams. My love for him permeates every corner of my soul, and as I look back with fond remembrance on our past adventures, I can’t wait to greet the future with the one person who truly understands me.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure

CHAPTER I

OVERVIEW

Cardiovascular disease (CVD) encompasses several disorders of the circulatory system that together serve as a leading cause of non-communicable disease death both in the United States and globally[1]. A common characteristic of CVD is thrombosis, and major thrombotic events are often associated with impaired fibrinolysis[2]. Fibrinolysis is mediated primarily by the interaction between tissue plasminogen activator (t-PA), which converts plasminogen to plasmin, and plasminogen activator inhibitor 1 (PAI-1), which inactivates t-PA[3]. Although both t-PA and PAI-1 are important for fibrinolytic balance, there has been more evidence that PAI-1 levels are correlated with CVD susceptibility, severity, and fibrinolytic potential[4,5]. PAI-1 is a critical modulator of fibrinolysis and greater understanding of its inter-individual variation may provide valuable information in the development of therapeutic intervention strategies for CVD patients. PAI-1 levels have been associated with increased risk/severity of CVD in multiple populations in several studies, but the underlying genetic architecture of PAI-1 variation remains poorly understood. While genetic studies have been performed to explain the variation in PAI-1 levels, the majority of these studies were performed in Caucasian and Asian cohorts. Even in the most assessed populations, studies have failed to explain most of the genetic effectors of PAI-1 even though heritability estimates are between 0.42 and 0.71[6,7]. African populations have some of the heaviest CVD disease burden worldwide, yet PAI-1 variation is poorly understood in these populations primarily because they have not been studied. The focus of this dissertation is the investigation of the genetic mechanisms involved in PAI-1 variation in both West African and Caucasian populations using single variant, gene-level, and pathway-level analyses. We hypothesize that the complex nature of PAI-1 variation will be better described through single and multivariate analysis at the single nucleotide polymorphism (SNP) level that acknowledges and accommodates for the skewed nature of the PAI-1 distribution as well as the likelihood of the existence of genetic variants with non-uniform and possibly non-linear effects. We further postulate that significant gene and biological pathway level effects may play a role in variation in PAI-1 levels.

The introduction and background of PAI-1 as a modulator of the fibrinolytic system, its role in the progression of thrombosis, and in the context of CVD risk and severity are presented in Chapter II, part A. In depth description of the study populations used in this dissertation as well as previous research performed in these cohorts are presented and the impact of these findings on the current study design are summarized in Chapter II, part B. Chapter II, part C summarizes preliminary research conducted in preparation to test the central hypotheses of the current study, as well as to compare and contrast previous strategies employed to investigate genetic variation in PAI-1 levels with those undertaken in this dissertation. Finally, Part D outlines the hypothesis and specific aims of this dissertation.

To explore common genetic variation effecting PAI-1 levels, quantitative regression based analyses are performed in the HeART cohort in Chapter III. Part A identifies common variants that effect PAI-1 levels across its entire distribution by examining the effect of genetic variation on median PAI-1 levels. These studies highlight novel associations between single variants in the arylsulfatase b (*ARSB*), carboxypeptidase A2 (*CPA2*), and leukocyte receptor cluster member 9 (*LENG9*) and median PAI-1 levels. Part B attempts to provide a more comprehensive look at PAI-1 variation by performing analyses directed at the upper quartile (75th percentile) of the PAI-1 distribution to uncover novel variants with significant impact on this clinically relevant portion of the PAI-1 phenotypic distribution. Upper quartile regression analyses led to the discovery of significant associations between single variants in exostosin glycosyltransferase 2 (*EXT2*) and period circadian clock 3 (*PER3*). The discovery of a significant association with *PER3* in the upper quartile supports previous evidence that the circadian pathway associates with regulation of PAI-1 levels in Caucasian populations as well as model organisms. This finding provides evidence that while the significance of the circadian pathway as a whole may be generalizable across Caucasian and African populations, the effect of specific genes may be population specific.

In Chapter IV, we present a novel approach, Multi-lOcus based selection of Candidate genes (MOCA) that integrates quantitative Multifactor Dimensionality Reduction q(MDR) analysis into a pipeline that exploits the information gained from multi-locus association analyses to prioritize gene regions likely to harbor functional variants associated with PAi-1 levels. MOCA operates under the premise that multi-locus effects play a significant role in variation of PAI-1 levels and incorporating multi-locus association analyses into the prioritization of genes for further evaluation following single locus analyses will increase our power to detect functional variants and enhance our ability to make direct biological/mechanistic connections between discovered variants and PAI-1. Our approach revealed four significantly associated multi-locus effects that localized to four loci located on chromosomes 5,8,17, and 20 identifying 28 novel candidate genes for PAI-1. Evaluation of the nature of the associated multi-locus models is presented in this chapter as well. Additionally, we highlight the utility of our approach by comparing MOCA identified regions of interest to previously discovered quantitative trail loci (QTL) for various CVD traits and/or biomarkers known to effect PAI-1 levels. We conclude Chapter IV with an example of the utility of the MOCA approach by presenting possible biological connections between the putative candidate genes identified by the region of interest on chromosome 5 and PAI-1.

CHAPTER II

INTRODUCTION

A. Introduction and Background

A1. Cardiovascular disease and thrombosis

Cardiovascular disease (CVD) does not describe a single condition, but refers to a number of different diseases that affect the peripheral blood vessels and/or the heart, with the most common form of CVD being ischemic heart disease. Two common, and sometimes severe endpoints of CVD, are stroke and myocardial infarction (MI). A stroke is an obstruction of blood flow to the brain, while MI is an obstruction of blood flow to the heart. Although sometimes used interchangeably, stroke and MI are two different phenomena that can both be characterized by thrombosis. Thrombosis, stemming from the Greek word *thrombos,* meaning "to clump", is defined as the formation of a blood clot in an undamaged blood vessel. There are two coagulation pathways that lead to thrombosis, the intrinsic and extrinsic coagulation pathways[3]. The intrinsic pathway is activated by trauma to the blood vessel which results in the exposure of plasma to a foreign substance; examples would be collagen from a damaged vascular wall or high molecular weight kininogen. The extrinsic pathway is activated by the exposure of plasma to tissue factor, an integral membrane protein found in non-vascular cells, after the vessel wall has been damaged. Though the pathways differ in their initiation, they converge at the production of tissue factor Xa, a key member of the prothrombinase complex. Prothrombinase converts prothrombin to thrombin, which is an essential step in clot formation. Thrombin then converts fibrinogen to fibrin. It is fibrin that creates the mesh that binds the blood clot. Clot formation plays an important and necessary role in minimizing blood loss after vascular injury.

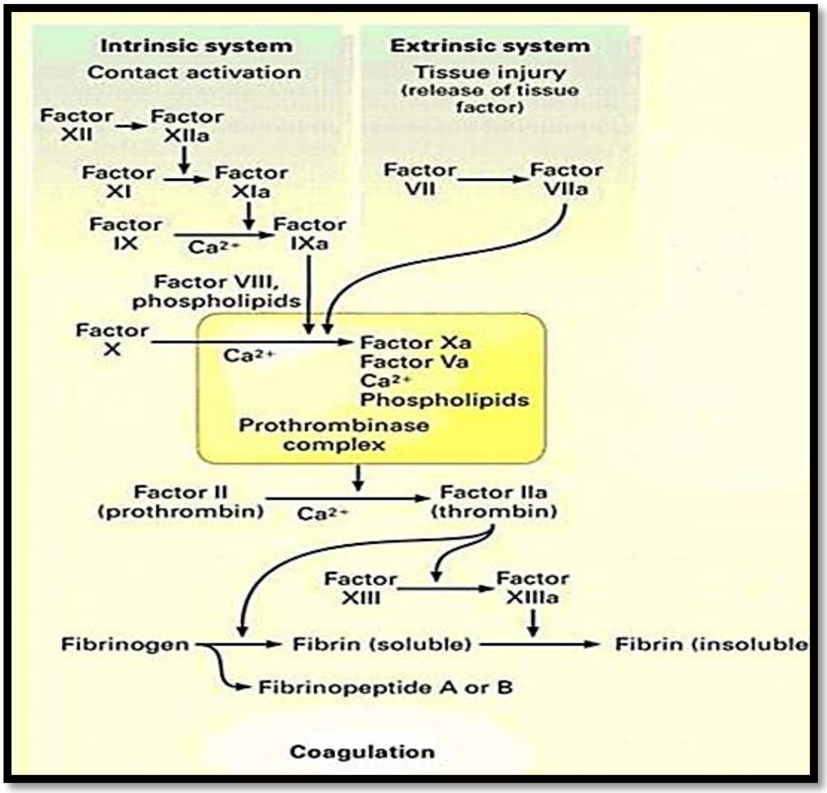


Figure 2-1. Intrinsic and Extrinsic Coagulation Pathways. Biomarkers involved in both intrinsic and extrinsic coagulation processes are presented. Picture adapted from Kohler *et al* 2000[8].

Under normal conditions, coagulation (clot formation) is carefully balanced by the process of fibrinolysis. Fibrinolysis is the process by which the insoluble fibrin mesh in blood clots is degraded by plasmin. Fibrinolysis stops clots from remaining in the circulatory system where they might cause harm to the body. Clots have the potential to break off and travel through the blood stream potentially causing damage by blocking blood vessels, a process known as embolism. Thrombosis occurs when there is an imbalance between the processes of coagulation and fibrinolysis; this could mean that blood clots are forming inappropriately or that the clots formed are not being properly degraded.

A2. Tissue Plasminogen activator (t-PA), Plasminogen Activator Inhibitor 1 (PAI-1) and Fibrinolysis

Major thrombotic events may be characterized by an imbalance in the relationship between clot formation and clot degradation. This imbalance could be instigated by a decrease in fibrinolysis. The fibrinolytic pathway is largely mediated by the relationship between tissue plasminogen activator (t-pa) and plasminogen activator inhibitor 1 (PAI-1)[3]. Fibrinolysis is initiated with the release of t-pa from endothelial cells following injury and in response to the presence of thrombin. T-PA then converts plasminogen to plasmin by cleaving the Arg561-Val562 peptide bond. Plasmin then degrades the fibrin clot, and any intact fibrinogen, into fibrin degradation products.

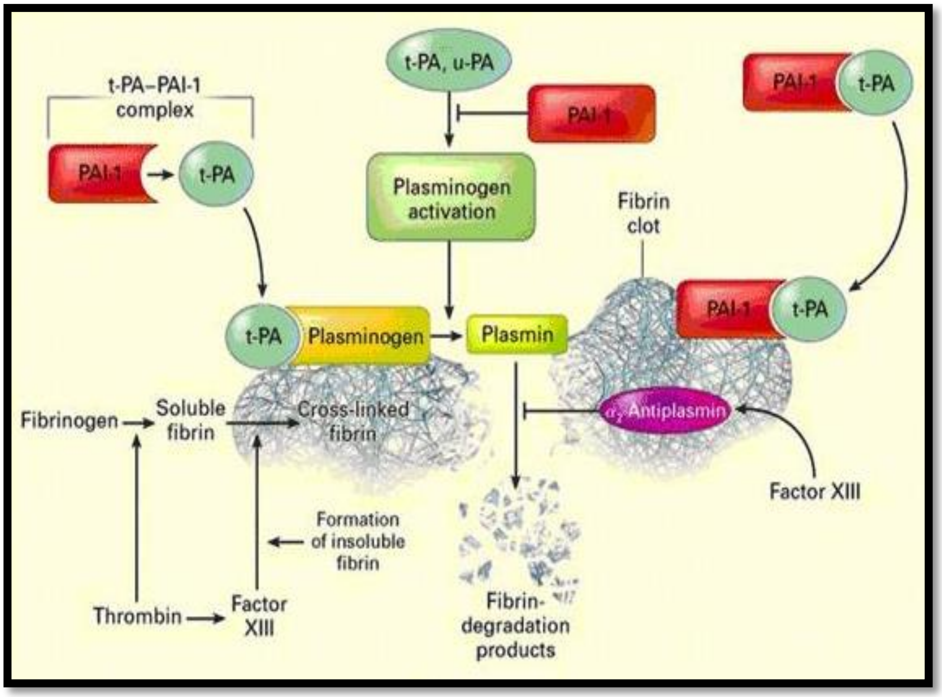


Figure 2-2. Key bioregulators of the fibrinolytic system. Summary of the action of t-PA and PAI-1 in the regulation of the fibrinolytic process. Picture retrieved from Kohler *et al* 2000[8].

The fibrinolytic pathway is regulated by PAI-1, which prevents excessive fibrinolysis by inhibiting t-PA. PAI-1 is an acute phase reactant protein synthesized by endothelial cells and hepatocytes that inhibits the action of t-PA by binding to the molecule and forming a complex[3]. Since PAI-1 is the principal inhibitor of t-pa, as well as urinary plasminogen activator, it plays an integral role in the regulation of the fibrinolytic pathway. Fibrinolysis is an important and dynamic process that performs the dual function of preventing the extension of clots beyond the site of injury and degrading fibrin clots that are no longer needed. It has been suggested that impaired fibrinolysis may be related to the clinical evolution of CVD, particularly coronary artery disease[9].

Due to their significant impact on the regulation of the fibrinolytic system, both plasma PAI-1 and t-PA are often used as biomarkers for fibrinolysis and as measures of fibrinolytic activity. It was discovered that increased plasma t-PA indicated an inhibition of fibrinolysis[10,11]. This may seem counterintuitive, as t-PA initiates fibrinolysis. This paradox is easily explained by the fact that free t-PA released into the blood rapidly forms a complex with circulating PAI-1 and these complexes have a longer half-life compared to free t-PA; therefore assays of t-PA antigen are mostly measuring inactive t-PA/PAI-1 complexes[12,13]. For the aforementioned reasons, increased plasma t-pa is highly correlated with increased plasma PAI-1 levels; in order to reduce redundancy in our analyses, this dissertation will focus exclusively on PAI-1 measurements as an indicator of fibrinolytic activity, and as a quantitative trait in all subsequent analyses.

A3. Plasminogen Activator Inhibitor – 1 and Cardiovascular Disease Risk and Severity

Studies have also shown PAI-1 levels to be related to adverse outcomes in CVD patients, as well as healthy individuals. High plasma PAI-1 levels have been shown in many studies to be associated with increased risk of CVD in the form of higher susceptibility to atherothrombotic diseases such as myocardial infarction and coronary artery disease[14-16]. One such prospective study, The Caerphilly Study, found baseline PAI-1 levels to be significantly associated with the occurrence of cardiovascular events, with a hazard ratio of 1.24 (95%CI:1.05 – 1.46, p=0.013)[14]. In a case control study involving 520 acute coronary syndrome cases, and 520 age and gender matched controls, plasma PAI-1 concentration (fifth quintile vs. the first quintile) was shown to be a significant (OR 5.3, 95%CI:1.2 – 23.8, p <0.05) and independent risk factor associated with the future occurrence of major adverse cardiac events[16]. Additionally, increased PAI-1 expression in human atherosclerotic lesions has been reported[17].

A4. Impact of genetic variation on Plasminogen Activator Inhibitor - 1 levels

There is clear evidence that genetic factors play a role in the variation of PAI-1 levels. Numerous twin studies have investigated the heritability of PAI-1, and although the exact estimates of heritability differ among the different studies, all reported heritability estimates have been significant. A twin study performed in a cohort consisting of 464 healthy twins estimated heritability of PAI-1 to be 0.44[18]. A Swedish study estimated the heritability of PAI-1 antigen to be 0.42[19], while yet another study lists PAI-1's heritability at 0.71[20]. It has also been shown in twin studies that there are genetic factors that influence variance in PAI-1 that are shared with those that influence BMI and triglycerides.

In addition to strong evidence from twin studies that genetic factors play a role in the variance of PAI-1 levels, a genetic variant in the PAI-1 promoter region, the 4G/5G variant, has been shown to be associated with plasma levels of PAI-1[21-23]. Additionally, in a meta-analysis comprising 18 studies performed by Tsantes *et al.*, the 4G allele was shown to be significantly associated with venous thrombosis (OR=1.15, 95%CI:1.33 – 2.54)[24]. Although the 4G/5G allele is the most studied, and validated, polymorphism effecting variation in the expression of PAI-1, it has been estimated to only explain 1-3% of the genetic variance[25-27]. This modest estimate suggests that the majority of the genetic factors that play a role in the variation of plasma PAI-1 are as yet undiscovered. Indeed, the variation in PAI-1 may be attributed to epistasis, genes in other pathways, or gene-environment interactions. Several candidate gene studies have reported genetic associations between variants in the renin-angiotensin and bradykinin systems that were significantly associated with plasma t-PA and/or PAI-1 levels[22,28-30]. After verifying the impact of the 4G/5G variant on PAI-1 levels, a recent meta-analysis identified three additional variants located in the aryl hydrocarbon receptor nuclear translocator-like (*ARNTL*), peroxisome proliferator-activated receptor gamma (*PPARG*), and intestinal mucin-like (*MUC3*) genes, respectively[31].

A5. Biological Effectors and Modifiers of PAI-1 levels

The etiology behind the expression of PAI-1 is complex, and variation in plasma PAI-1 levels is likely to be influenced by numerous genetic and environmental factors. There are several environmental/biological factors that have been shown to influence PAI-1 levels; among these are alcohol use, body size, gender, the presence of insulin resistance, and triglycerides (Table 2-1).

Table 2-1. Biological and Environmental Effectors/Modifiers of PAI-1

Biological /Environmental Effector of PAI-1	Nature of Impact on PAI-1	Evidence
Alcohol consumption	Increased alcohol consumption associates with increased PAI-1	Djousse, L. *et al.* 2000[30], Pieters, M. *et al.* 2010[36], Mukamal, K. *et al.* 2001[26]
Body Size (obesity / BMI)	Increased PAI-1 levels correlate with obesity	Appel, S. *et al.* 2005[27]; Kenny, S. *et al.* 2013[28], Berberoglu, M *et al.* 2006[29]
Insulin Resistance	Increased PAI-1 levels are a hallmark of disease in insulin resistant patients	Bastard, J. *et al.* 2000[38], Ridker, P. *et al.* 2004[7]; Juhan-Vague, I. *et al.* 2000[40]
Gender	Gender-specific effects have been discovered between genetic variants and PAI-1 levels; gender impacts PAI-1 correlations with CVD biomarkers	Asselbergs, F. *et al.* 2006[49], Harslund, J. 2007[48]; Asselbergs, F. *et al.* 2007[50]
Age	Increasing age is correlated with increased PAI-1 expression	Sobel, B. *et al.* 2006[46]; Cesari, M. *et al.* 2010[45]; Ardite, E. *et al.* 2012[47]
Triglycerides	Triglycerides stimulate PAI-1 production, and may modify the impact of other factors on PAI-1 expression	Pieters, M. *et al.* 2010[36]; Ma, L. *et al.* 2004[35]; Brown, N *et al.* 2001[41]

Alcohol consumption and obesity are two biological factors that have been shown in humans and model organisms to effect PAI-1 levels[32-36]. Heavy alcohol consumption, defined as greater than seven alcoholic beverages per week, is associated with higher levels of PAI-1; it has been postulated that alcohol may lower the fibrinolytic potential through the activation of RAS either as a direct effect of acetaldehyde or by the inhibition of vasopressin release caused by volume depletion[32,37,38]. Obesity has been shown to be associated with increased PAI-1 production, as fat cells are a major source of PAI-1[39]. Interestingly, it seems that the distribution of fat plays a critical role in this increase as subcutaneous adipose tissue produces less PAI-1 than visceral adipose tissue. A study by Sartori et. al cited that the 5' 4G/5G variant associated with PAI-1 levels in patients with central, but not peripheral, obesity[40]. Although several studies have shown a correlation between PAI-1 levels and obesity, it still remains unclear whether PAI-I levels increase in response to the presence of obesity, or if PAI-1 is causative[41].

Similar to obesity and alcohol consumption, triglycerides and insulin resistance have also been shown to be correlated with or modify the expression of PAI-1 levels in both model organisms and humans[41-48]. Triglyceride rich lipoproteins have been shown to stimulate PAI-1 production *in vivo*, and PAI-1 has been found to correlate with increased triglycerides in both mice and humans[41,42,49]. Other studies revealed that the correlation between PAI-1 and triglyceride levels is heavily influenced by PAI-1 4G/5G variant genotype[47]. Triglyceride levels have also been shown to influence the relationship between alcohol consumption and PAI-1[42]. Elevated PAI-1 levels are a hallmark of disease etiology in insulin resistant patients, and insulin resistance has long been considered to be a major modulator of PAI-1 expression; it has been suggested that PAI-1 may be the common link between obesity, insulin resistance and cardiovascular disease[50].

Age and gender associate with and/or modify associations between PAI-1 and other biological and/or genetic factors[28,51-57]. Pai-1 has been hypothesized to contribute to CVD in the elderly via the age-related development of a prothrombic state in the fibrinolytic system; further research in model organisms has shown that PAI-1 expression in the heart increases as a function of time independent of insulin resistance [51,52]. Gender is a potent modifier of the relationship between PAI-1 and genetic and biological variants[55,56]. The biological factors mentioned in this subsection, as well as other common CVD risk factors such as BMI, cholesterol and diet, may modify or confound the effect of our genetic variants on plasma PAI-1 levels, if their influence on PAI-1 is not addressed. We have therefore adjusted our analyses to account for these factors to the extent that we are able.

B. Description of Study Population and Previous Studies

B1. Description of HeART Cohort

The Hypertension and ARterial Thrombosis (HeART) study is a population based cohort of 3431 subjects, consisting of 2375 urban and 1056 rural residents, recruited from the Brong Ahafo regional capital of Sunyani, Ghana between 2002-2003[58]. To date, the HeART cohort represents one of the largest population-based studies of inter-individual variation in plasma PAI-1 levels for any ethnic group in sub-Saharan Africa. The cohort is approximately 60% female and includes individuals between the ages of 18 and 80. Recruitment for the cohort was based on word-of-mouth, without regard to chronic disease status, and includes only unrelated individuals. Subjects were excluded from the study if they showed signs of acute illness, were less than 18 years of age, or were first or second degree relatives of someone already enrolled in the study. Morning blood samples were collected from all participants. Demographic and medical information was also collected, as well as multiple measures of CVD risk, including BMI, fasting lipids, and fasting glucose. A subset of 992 urban individuals with PAI-1 measurements were previously selected for candidate gene analyses aimed at investigating the genetic effects of variants in the RAS on variation in PAI-1 levels[28,57]. For this dissertation, a subset of 1152 urban individuals from the HeART cohort, which included the 992 previously genotyped individuals, was selected as our study population.

B2. Previous Studies of PAI-1 in the HeART Cohorts

B2a. Gender specific effects of polymorphisms on plasma t-PA and PAI-1 levels

Previous studies performed in the HeART cohort found that gender affected the association of polymorphisms from genes in the renin-angiotensin and fibrinolytic systems, as well as polymorphisms in the PAI-1 gene[59]. After adjustment for non-genetic factors, the most significant genetic association with PAI-1 levels in males was with the rs4646972(*TPA* I/D polymorphism) (p=0.014). Alternatively, in females rs1799768 (PAI-1 4G/5G promoter polymorphism) was the most significant SNP (p=0.001). Tests of homogeneity of regression also showed differential effects of BMI (p=0.032) and triglycerides (p=0.011) on plasma PAI-1 levels between males and females, further emphasizing the complexity of PAI-1 levels as a phenotype and highlighting the universal association modifying effects of gender on variation in PAI-1 levels.

B2b. Epistatic effects of polymorphisms on plasma PAI-1 levels

Previous studies in the HeART cohort investigating epistatic interactions between polymorphisms in the renin-angiotensin (REN) and fibrinolytic systems[28]. This study found the most significant interactions associated with plasma PAI-1 levels to be rs1464816 (*REN* G6567T) and rs4646972 (*TPA* I/D) in males (p=0.032), and rs3730103 (*REN* T9435C) and rs4646972 (*TPA* I/D) in females (p=0.001). The two locus interaction terms explained as much as 2.6% of the variance in PAI-1 protein levels, measured by r^2, as compared to previous single SNP analyses that explained less than 1.3% of the variance[59]. These results provided evidence of significant genetic effects that would not have been detected using single SNP analyses alone.

C. Hypothesis and Specific Aims

Hypothesis: **Common genetic variants are associated with variation in plasma PAI-1 levels, and these effects can be uncovered via analyses at the single nucleotide polymorphism (SNP), multi-locus (SNP-SNP interactions) and gene levels.**

Specific Aim I. Identify significant associations between common SNPs and PAI-1 levels in the HeART Cohort.

a. *Identify significant single locus effects on median PAI-1 levels*
b. *Identify significant single locus effects on the upper quartile of PAI-1 levels*

The Illumina Infinium HumanExome chip will be supplemented with ~10,000 common SNPs chosen from based on putative association with PAI-1 variation, and used to genotype 1152 individuals from the HeART cohort. Quantile regression analyses will be performed in the HeART cohort to identify significant genetic effects on the median and upper quartile of PAI-1 measurements. Ordinary Least Squares (OLS) regression analyses will be performed on SNPs showing evidence of significant effects identified through median regression analyses, and the sensitivity and specificity of the two tests will be compared. All regression analyses will be adjusted for gender, BMI, age, triglycerides, and genotype at the PAI-1 4G/5G promoter variant. Bonferroni correction will be used to correct for multiple testing bias. SNPInfo, a comprehensive bioinformatic web-based algorithm, will be used to predict the possible functional impact of associated SNPs.

Specific Aim II. Presentation of a Novel Approach to incorporate multi-locus association signals into selection of Candidate genes for Prioritization in Future Studies

a. *Presentation and application of a novel approach, MOCA, to identify regions likely to harbor functional variants associated with phenotype*

1. *Evaluate and characterize significant intragenic interactions discovered as a result of the MOCA approach*

We present MOCA (Multi-lOcus based selection of Candidate genes), an novel approach that integrates quantitative multifactor dimensionality reduction (qMDR) analyses into a pipeline that exploits multi-locus association effects to identify genic regions predicted to harbor candidate genes with functional effects on PAI-1. MOCA operates under the premise that multi-locus association signals will “tag” candidate genes for prioritization more effectively than single variant signals. By incorporating qMDR analysis into our approach to generate multi-locus association signals, which we then assign to gene regions, we will be able to localize association signals to regions that encompass a few genes which we will then select to include in further analyses of PAI-1 as a necessary next step after single variant level analyses. Although the main aim of MOCA is to identify regions of interest to identify candidate genes, this approach also allows for the possibility of identifying significantly associating true (synergistic) epistatic effects. In the event that such effects are identified, we will evaluate and characterize these effects. This approach will provide novel information about the nature of the underlying genetic architecture of PAI-1 that will be used to further elucidate genetic factors affecting variation in this potent CVD biomarker.

CHAPTER III

EXAMINING SINGLE VARIANT EFFECTS ON MEDIAN AND ELEVEATED PAI-1 LEVELS IN A GHANIAN COHORT

Overview parts A and B

PAI-1 is a major modulator of the fibrinolytic system, and is associated with cardiovascular disease (CVD) susceptibility and severity. Although heritability estimates for PAI-1 are large, the few genes associated with it explain only a small portion of inter-individual phenotypic variation. Additionally, most genetic studies of PAI-1 levels have been performed in European descent populations, and the few studies that have been performed in African populations assessed only a small number of variants. Almost all genetic studies have employed ordinary least squares (OLS) regression to assess the impact of genetic factors on mean PAI-1 levels. For studies of PAI-1 it has been standard practice to log-transform PAI-1 and employ OLS regression models to detect significant effects even when transforming the data does not normalize the distribution. In extremely large sample sizes (e.g., those used for meta-analyses) strong effects are still detectable using this approach. However, most single cohort genetic studies have moderate sample sizes, and it is probable that most single variants will have small to moderate effects on PAI-1 variation making OLS less than an ideal analytical strategy.

In Chapter III, to address the inferential limitations of previous studies, non-parametric methods were used to evaluate the effect of approximately 39,000 SNPs on PAI-1 levels in a Ghanaian cohort. Part A investigates SNP effects on median PAI-1 levels using median regression, a non-parametric alternative to OLS. We also reassessed our significant associations using OLS to determine if violation of the normality assumption impacts findings in our data. Significant associations between non-synonymous SNPs and median PAI-1 were detected in *arylsulfatase B*

(ARSB), *carboxypeptidase A2 (CPA2)*, and *leukocyte receptor cluster member 9 (LENG9)*. We found that although both median regression and OLS returned comparable effect sizes, OLS models had universally higher standard error rates and larger p-values due to misspecification of the model; if inference had been based on OLS regression analyses our associated variants would not have been detected after correction for multiple testing.

Elevated PAI-1 levels, in particular, have been shown to associate not only with susceptibility to CVD, but also with increased disease burden via recurrent adverse events. Part B assesses the impact of SNP genotype on the upper quartile of PAI-1 distribution using quantile regression, a model-free alternative to OLS that does not assume a uniform effect of SNPs across the phenotypic distribution. Nineteen SNPs were significantly associated with variation in the upper quartile. Of particular note was an association with rs10462021 ($p= 2.07 \times 10^{-6}$), a missense variant located in *period circadian clock 3* (*PER3*) because a recent meta-analysis in Caucasian populations identified another circadian clock gene, *aryl hydrocarbon receptor nuclear translocatior-line gene* (*ARNTL*) significantly associated with PAI-1 levels. There was no overlap between the variants that were significantly associated with median PAI-1 and those associated with the upper quartile of the distribution, revealing for the first time to our knowledge, the presence of quantile-specific effects on PAI-1 in an African cohort.

These studies revealed novel associations with median and elevated PAI-1 levels , and provide evidence for the generalizability of the circadian pathway's effect on PAI-1 levels. They also highlighted the utility of employing non-parametric methods when standard analytical methods are not appropriate for a particular study, and demonstrated the possible impact of model misspecification on statistical inference. The latter may, at least partially, explain why moderately sized studies have been unable to identify the genetic variants responsible for the majority of the genetic impact on PAI-1 variation.

A. Genetic Impact of common SNPs on Median PAI-1 levels in the HeART cohort

Introduction

Cardiovascular disease (CVD) describes multiple conditions of the circulatory system that overlap in terms of environmental and genetic risk factors, symptoms, and disease etiology. It is responsible for approximately 48% of all non-communicable disease (NCD) related deaths worldwide[1]. Thrombosis is a major factor in the etiologies of several cardiovascular diseases, including myocardial infarction (MI) and stroke, and represents an excellent target for the prevention and treatment of this disease burden[60,61]. Fibrinolysis, the dynamic process by which fibrin is cleaved by plasmin to promote clot degradation and inhibit thrombus formation, mediates thrombotic events[62]. The impairment of the normal fibrinolytic balance, due, at least in part, to increased plasminogen activator inhibitor-1 (PAI-1) levels predicts thrombotic risk and severity[63]. Although several studies of plasma PAI-1 levels indicate a positive correlation between increased PAI-1 and increased susceptibility to thromboembolism, atherosclerosis, myocardial infarction, and by proxy, most CVDs, the nature of the relationship between PAI-1 and CVD risk remains controversial, and as yet undefined[5,64-76].

PAI-1 is a biomarker of fibrinolytic activity, and its levels are heavily influenced by genetic variation[6,7,77,78]. The most studied, and to date, strongest single genetic variant impacting PAI-1 levels is the 4G/5G promoter polymorphism which has been shown in several studies to influence circulating PAI-1 levels in a dose dependent manner with carriers of the 4G allele exhibiting higher levels of mean circulating PAI-1[79]. However, this variant alone does not account for all variation in PAI-1 levels attributed to genetic factors, and there is evidence that other variants, either in PAI-1 or in other genes, with more moderate effect sizes also play a significant role in the variation of PAI-1 levels[79-81]. The majority of studies aimed at uncovering the genetic factors affecting PAI-1 have been conducted in Caucasian and/or Asian populations, including several recent meta-analyses and GWAS studies[31,82,83]. A few studies have been performed in African populations as well, most notably using individuals from the HeART Cohort, a population-based cohort from the Brong Ahafo region in Sunyani, Ghana[28,48,57]. The African based studies, however, were candidate gene analyses focused on the individual or combined impact of only a few single nucleotide polymorphisms (SNPs) selected based on *a priori* biological knowledge. This approach ensures focused attention on markers that are likely to have a significant effect on phenotype, but is unable to identify novel variants not previously known to affect PAI-1 variation.

Another common, and possibly limiting, factor in all previously performed analyses is the use of standard linear regression, or ordinary least squares (OLS) regression, to determine the “overall” impact of SNP genotype on PAI-1 levels. This approach, though standard in genetic association studies, may not be the most appropriate analysis due to the highly skewed nature of the PAI-1 distribution. The non-normality of the PAI-1 distribution in most analyzed cohorts also violates the assumption of normality required of standard linear regression[84]. Furthermore, the use of the mean as a measure of the overall behavior of the dependent variable in the presence of extreme outliers, another characteristic of the PAI-1 distribution, is also problematic as this may lead to incorrect inference[85]. A more appropriate measure of the overall behavior of the PAI-1 distribution in response to SNP genotype, is the median of the distribution. Median regression is a non-parametric method that parallels OLS regression, but is robust to non-normality, heteroskedasticity, and outliers in the phenotypic distribution, and measures the effect of SNP genotype on the median of the dependent variable as opposed to mean [84,86,87].

In this report, we evaluated the effect of SNP genotype on variation in PAI-1 levels using median regression analyses in a Ghanaian cohort genotyped for approximately 39,000 common SNPs. We also explicitly tested the effect of violation of OLS assumptions and model misspecification on statistical inference by performing complementary OLS regression analyses of variants significantly associating with PAI-1 in median regression analyses. We also employed bioinformatic/data mining techniques to assess the putative functional impact of significant SNPs and present possible biological connections between significant loci and PAI-1 levels.

Materials and Methods

Subjects

The Hypertension and ARterial Thrombosis (HeART) cohort consists of approximately 3400 urban and rural residents of the Brong Ahafo regional capital of Sunyani, Ghana recruited between 2002 and 2005, is one of the largest collections of PAI-1 measurement and lipid biomarker in a West African population, to date. HeART cohort ascertainment, DNA collection, biomarker measurement protocols, inclusion, and exclusion criteria have been previously described in detail elsewhere[48]. Our study cohort consisted of 1105 unrelated urban individuals selected from the HeART cohort. Selection criteria for our cohort included 992 individuals who had been previously genotyped and assessed in PAI-1 genetic studies[28,57] (n= 992) as well as an additional 113 subjects from the 90th percentile of the plasma PAI-1 distribution (urban residents only) of the total HeART cohort. Inclusion criteria for the study cohort were the availability of a viable DNA sample and demographic /clinical data including age, body mass index (BMI), triglycerides and PAI-1 measurements; clinical and demographic data was directly measured as described by Williams et al[48].

Genotyping Scheme

DNA from the study cohort was genotyped using the Illumina Infinium HumanExome BeadChip (Exome Chip) platform (Illumina, Inc., San Diego, CA). The Exome Chip provides focused coverage of the exonic regions of the genome using approximately 240,000 markers. This coverage was further supplemented by 8,439 common variants selected to provide focused coverage of target genes with previous evidence of association with variation in PAI-1 levels.

Quality Control Procedures

PAI-1 measurement Quality Control Criteria

As mentioned above specific protocols for biomarker measurement have been described elsewhere, protocols specific to PAI-1 measurement will be briefly summarized here. Due to the circadian rhythm defined variability in PAI-1 levels, PAI-1 concentrations were measured from blood samples collected between 8:00 and 10:00 AM local time. PAI-l levels were measured using a commercially available enzyme-linked immunoassay (TriniLIZE PAI-1 Antigen Assay, Tcoag, Bray, Ireland). The sensitivity of this assay is 0.5 ng/ml PAI-1; any measurements below this threshold were converted to 0.25 ng/ml PAI-1 (0.5 x test detection limit) as per the suggestion of the assay manufacturer. Duplicates were randomly assessed for quality assurance; for instances in which duplicate readings were concordant (within the same quintile of the PAI-1 distribution) an average value was determined for the specified sample. Non-concordant duplicate samples were reassessed; if concordance was not reached upon reassessment then the sample was removed from the study.

Study Participants and SNP Quality Control Criteria

Study participants were evaluated for genotyping efficiency and completeness of demographic data and biomarker measurements. Subjects with genotyping efficiency less than 95% and/or missing demographic and/or biomarker data were excluded from the study. After quality control procedures, 1053 individuals (441 males, 612 females) remained. A total of 39,124 common variants (minor allele frequency (MAF) ≥ 0.05) markers were extracted from the Exome Chip genotype data. Quality control criteria for the selected common variants included genotyping efficiency ≥ 95% and Hardy-Weinberg equilibrium p value < 0.001. After quality control procedures were completed, 38,871 variants remained for inclusion in downstream analyses. All quality control procedures were carried out using the PLINK software package[88]. MAF and HWE p-values are presented for SNPs found to be significantly associated with median PAI-1 levels in Appendix Table 1.

Statistical Analysis

Preliminary Analyses

Family data was collected and utilized as a part of the inclusion criteria for all participants of the HeART cohort (individuals had to report that both parents and both sets of grandparents were native to Ghana), making the probability of population stratification on the basis of hidden substructure due to admixture low. However, as an added precaution, we explicitly tested for the presence of population stratification within our dataset using the STRUCTURE software program[89]. STRUCTURE analysis was performed using 8521 common variants genotyped in the HeART participants as well as the JPT+CHB, YRI, and CEU HapMap populations. STRUCTURE runs used an admixture model with correlated allele frequencies with a burn-in period of 8,000 runs with 500 subsequent iterations. Individuals from the HapMap populations were assigned to their corresponding predetermined clusters and used as founder populations in the STRUCTURE analysis to detect possible admixture or misclassification of HeART participants. As expected, STRUCTURE analysis revealed no significant evidence of population stratification within our dataset Appendix Figure 1.

Prior to association testing, the distributions of demographic and biological variables were assessed in males and females, separately, to determine if any significant differences existed between genders. These comparisons are presented in Table 3-1. Normality of continuous traits was evaluated using the Shapiro-Wilkes test ($p < 0.05$). For normally distributed continuous variables (Shapiro-Wilkes test $p > 0.05$) the Student's t test was used to assess mean differences between genders. In cases of non-normality (Shapiro-Wilkes test $p < 0.05$) the Wilcoxon rank sum test was used. For discrete variables, such as genotype at the *PAI-1* 4G/5G promoter variant, the Chi-square test was used to determine if significant differences existed between genders. All aforementioned preliminary assessments were performed using the STATA 11 statistical package[90].

Table 3-1. Gender-separated demographic and clinical characteristics.

		Males (n=441)	Females (n=612)	P-Value[1]
Age (years)		44.02(12.47)	43.22(10.75)	0.523
Body Mass Index (kg/m^2)*		24.21(4.29)	27.08(5.44)	<0.001
Triglycerides (mg/dL)*		94.45(52.85)	93.89(56.24)	0.489
serum PAI-1 levels (ng/mL)*		7.96(8.90)	8.84(11.02)	0.930
PAI-1 4G/5G genotype	4G/4G	20	34	0.264[2]
	4G/5G	108	171	
	5G/5G	260	332	

*; mean (standard deviation) untransformed variables are presented

[1] P-values are from the Wilcoxon Rank Sum test unless otherwise indicated.

[2] P-values are derived from the Chi-square test of association.

Prior to association testing, the distribution of PAI-1 levels were assessed for normality and found to be non-normal (Shapiro-Wilkes test $p < 0.001$). Non-normality of the distribution persisted after the data was log-transformed. Graphical presentation of the nature of the distribution of PAI-1 measurements (ng/ml) before and after log-transformation in our study cohort is presented in Figures 3-1a. and 3-1b.

Figure 3-1a. Nature of the PAI-1 Distribution in the HeART Cohort.

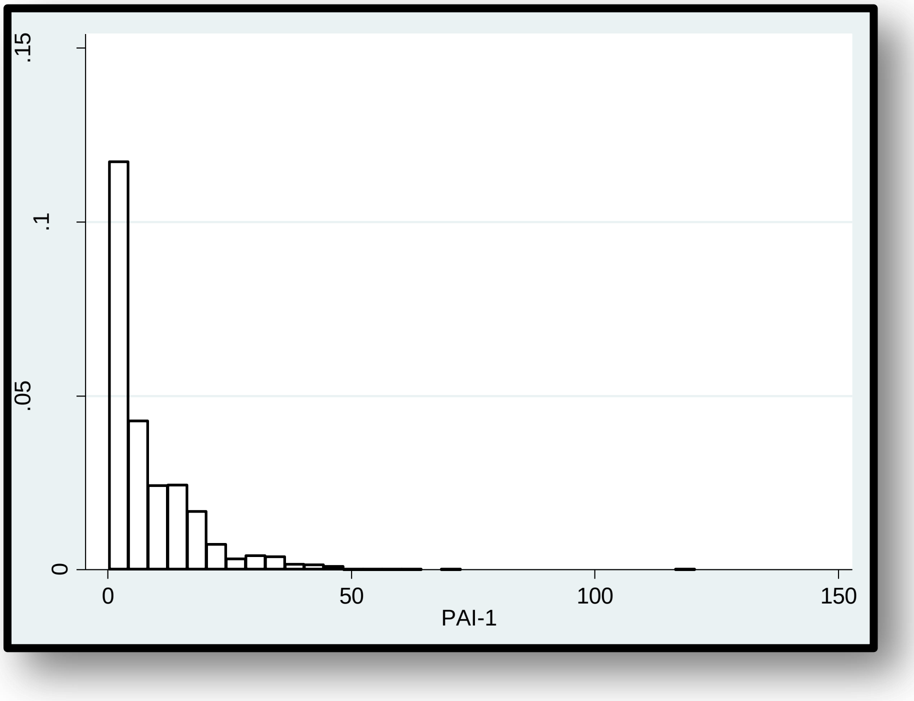
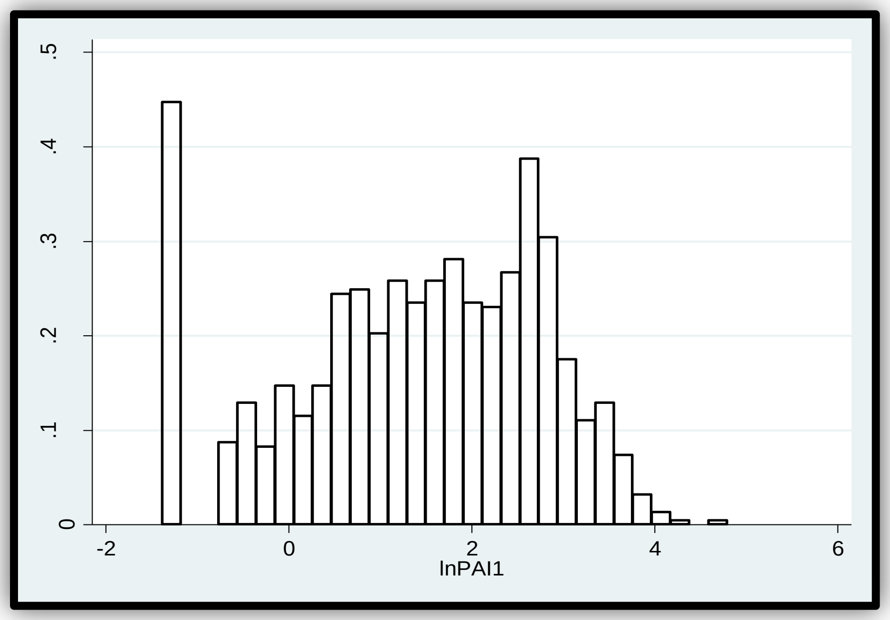


Figure 3-1b. Nature of the PAI-1 Distribution in the HeART Cohort after Log-Transformation.



Although log-transformation was unable to normalize the PAI-1 distribution in our cohort, log transformed PAI-1 measurements were still used as the dependent variable in our analyses for two reasons 1.) to maintain the standard currently employed in the field, and 2.) because the non-parametric median regression analyses that will be used for association analyses is invariant to monotonic transformations (i.e. logarithmic transformation)[85].

Median Regression Analysis

The non-normal and heavily skewed nature of the PAI-1 distribution in our cohort, which persisted after log-transformation, violates the assumption of normality characteristic of OLS regression[85]. This violation can cause model estimates to be distorted and ultimately lead to incorrect model inference. Therefore, median regression was used as an alternative. Median regression is a non-parametric regression method that is robust to deviations from normality and the presence of extreme values and heteroskedasticity in the data[84]. Generally defined, there are two main differences between linear regression and median regression. Linear regression uses linear methods to minimize the sum of the squared error and describes the impact of the independent variable(s) on the mean of the dependent variable, while median regression minimizes the sum of the absolute value of the error term and assesses the median of the dependent variable[87].

To investigate significant associations between common variants from the ExomeChip and log-transformed PAI-1 levels, median regression analyses were performed using the quantreg package in the R software suite[91]. Regression models were adjusted for age, gender, BMI, triglycerides, and genotype at the *PAI-1* 4G/5G variant. Triglyceride levels were log transformed prior to model inclusion because the raw measures were not normally distributed (Shapiro-Wilkes test $p < 0.05$). Single variant results were visualized using Manhattan plots created using the qqplots package in R[92]. In instances where a significant association was found between single variants and PAI-1, enhanced images of the regions proximal to the associating variants were generated using the LocusZoom online tool[93]. A significant number of variants in our dataset were in moderate to high linkage disequilibrium (LD), violating the assumption of independence employed by the standard Bonferroni correction used to control for multiple testing; this would have made the test overly conservative in our dataset. As an alternative, False Discovery Rate (FDR) was used to correct for multiple testing, and a q of 0.1 was used to determine the FDR threshold value[94]. A total of 38,871 single SNP association tests were performed; with q=0.1 the FDR significance threshold for this magnitude of tests is $p< 2.57x10^{-6}$.

Median regression has been shown to exhibit greater sensitivity in detecting linear and non-linear effects than its parametric counterpart, OLS regression; particularly in instances of smaller effect sizes[85]. In order to explicitly test this phenomenon in our study, we constructed OLS regression models to evaluate the reported associations for SNPs that were revealed to be significant by median regression. OLS regression models were adjusted for age, gender, BMI, triglycerides, and genotype at the PAI-1 4G/5G variant. OLS regression analyses were performed in STATA11[90].

SNPs were coded additively to test the effect of the minor allele for all regression models. In cases where there were fewer than five individuals in a genotype group, SNPs were coded dominantly for the effect of the minor allele; the homozygous minor and heterozygote genotype groups were combined into one class and compared to the homozygous major genotype. SNP coding and genotypic distributions are outlined in further detail in Appendix Table 2.

Bioinformatic / Data Mining Investigation of Associating markers

Information from bioinformatic / data mining analyses aimed at determining possible functional consequences associated with significant variants was assessed with SNPinfo[95]. SNPinfo is a comprehensive web-based database that incorporates several independent algorithms to provide functional predictions for specified genetic variants. Specifically, the Function SNP Prediction (FuncPred) pipeline in SNPinfo incorporates the usage of several software tools/web servers such as

PolyPhen[96], SNPs3D[97], MATCH[98], and ESEfinder[99] to assess the possibility that assessed SNPs may affect biological function (i.e. protein structure/stability, exon splicing, transcription, etc.).

Results

Median Regression Analyses

Three non-synonymous single nucleotide polymorphisms (SNPs) remained significantly associated with circulating PAI-1 levels after both FDR and Bonferroni correction. These were rs1071598 ($p = 1.09 \times 10^{-7}$), rs61997065 ($p = 3.56 \times 10^{-7}$), and rs10406453 ($p=2.58 \times 10^{-7}$), located on chromosomes 5, 7, and 19, respectively (Table 3-2, Figure 3-2). Of the three significantly associated variants, rs1071598, a missense SNP located in the *arylsulfatase B* (*ARSB)* gene, and rs10406543, a missense variant in *leukocyte receptor cluster member 9* (*LENG9)* displayed a similar trend on median PAI-1 levels (rs1071598 $\beta = -0.442$, rs10406543 $\beta = -0.467$) (Table 3-3). In contrast rs61997065, in *carboxypeptidase A2* (*CPA2),* displayed a strong positive effect on median PAI-1 levels associated with increased copies of the minor allele ($\beta = 0.503$) (Table 3-3). With the exception of rs10406453 on LENG8, none of the three significantly associated variants were in high LD with nearby markers (Figures 3-3a-c).

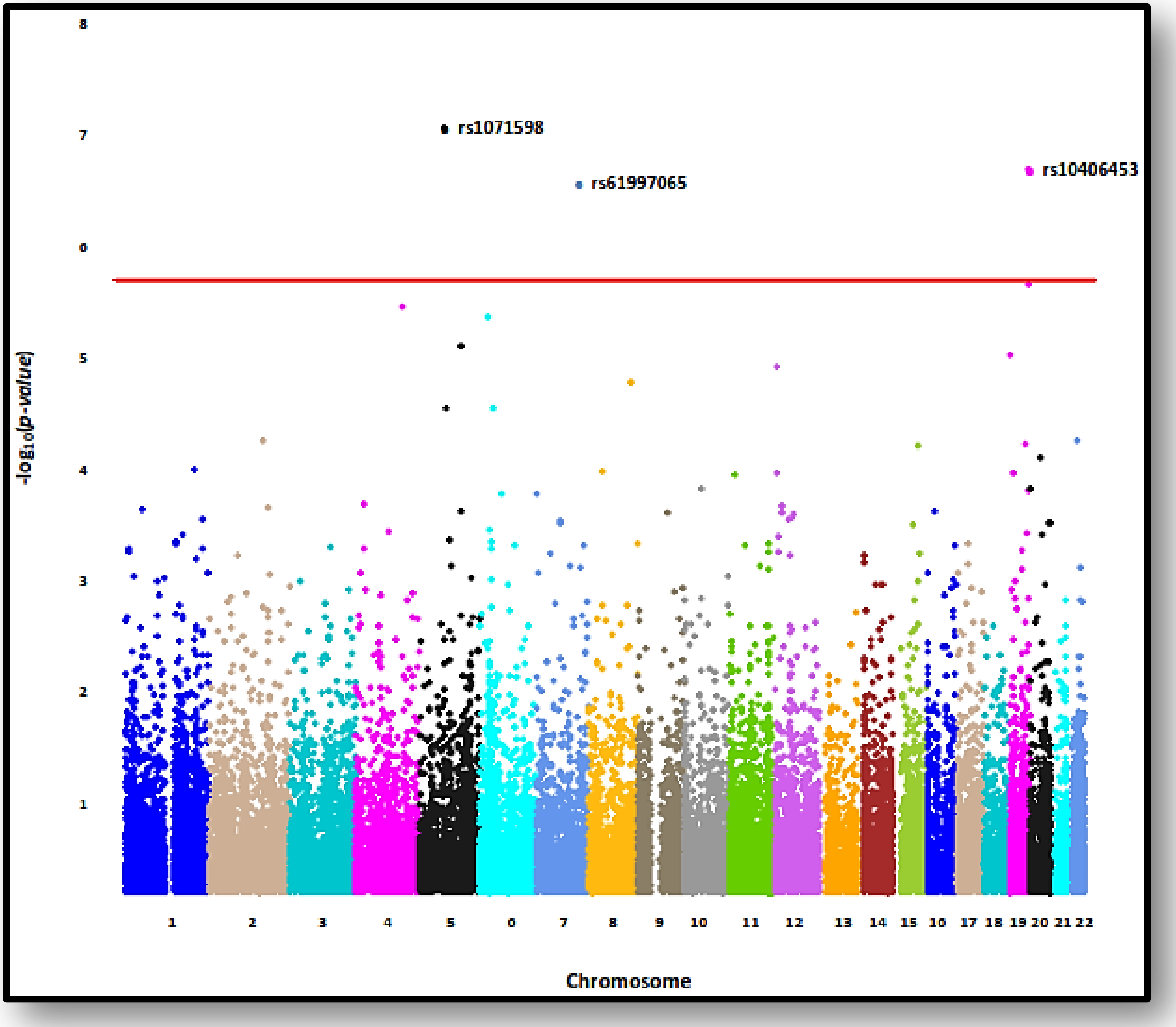
Table 3-2. Median Regression Results for Single Variant association with plasma PAI-1 levels

Chr.	Gene	SNP	Beta[1]	SE[2]	95% Confidence Interval		P-value
					LL	UL	
4	*SLC7A11*	rs4479754	-0.293	0.064	-0.418	-0.168	4.67E-06
5	*ARSB*	rs1071598	-0.442	0.083	-0.604	-0.280	**1.09E-07***
7	*CPA2*	rs61997065	0.503	0.098	0.311	0.695	**3.56E-07***
19	*LENG9*	rs10406453	-0.467	0.090	-0.643	-0.290	**2.58E-07***
19	*LENG8*	rs1035451	-0.375	0.080	-0.532	-0.219	2.90E-06

Results that remained significant after FDR correction are highlighted in **bold**; only a subset of significant results are presented above (p-value ≤ 10^{-5})

[1] Beta coefficient from median regression model represents the effect of the minor allele; model covariates: age, gender, BMI, triglycerides, and PAI1 4G/5G variant genotype.
[2] SE; Standard error; robust standard errors are reported above.
LL= 95% Confidence Interval lower limit; UL= 95% Confidence Interval upper limit
*****Effect remained significant after Bonferroni correction (threshold = 1.29E-06)

Figure 3-2. Manhattan Plot of SNP Association Analysis with Median plasma PAI-1 Levels.



Note: Only markers on chromosomes 1-22 are presented above; regions from the X chromosome, Y chromosome, pseudo-autosomal region of the X chromosome, and mitochondrial markers have been excluded. Statistically significant markers are labeled in bold. Red Line represents FDR significance threshold (2.57×10^{-6})

Figure 3-3a. LocusZoom visualization of the region proximal to rs1071598 located in *ARSB*

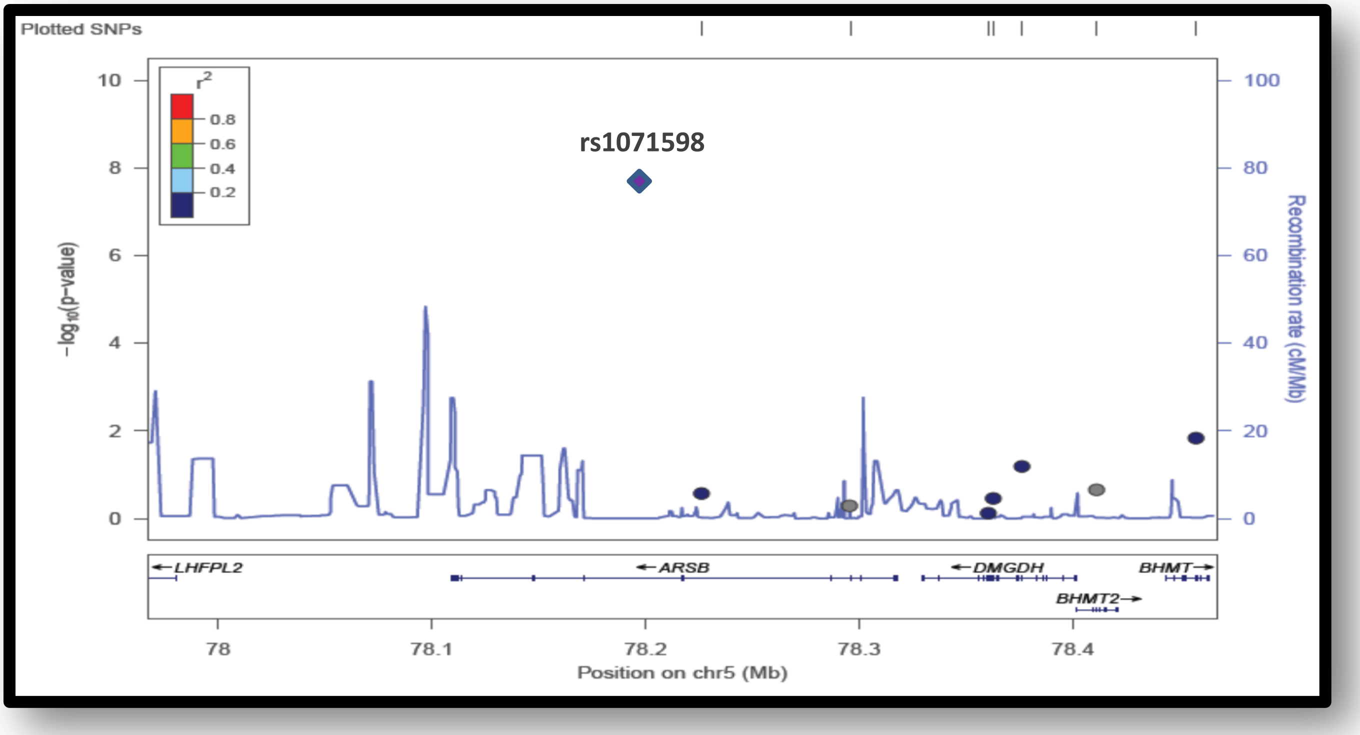


Figure 3-3b. LocusZoom visualization of the region proximal to rs61997065 located in *CPA2*

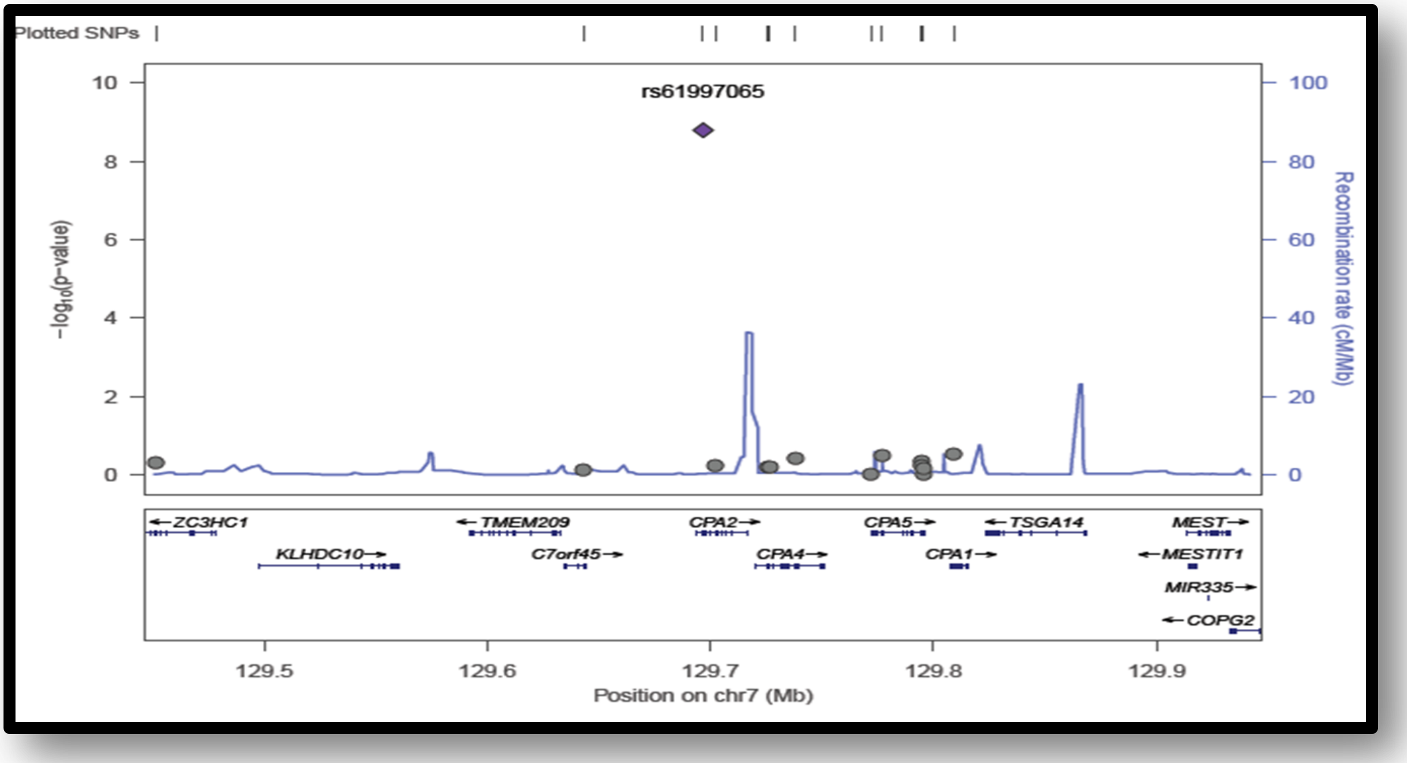
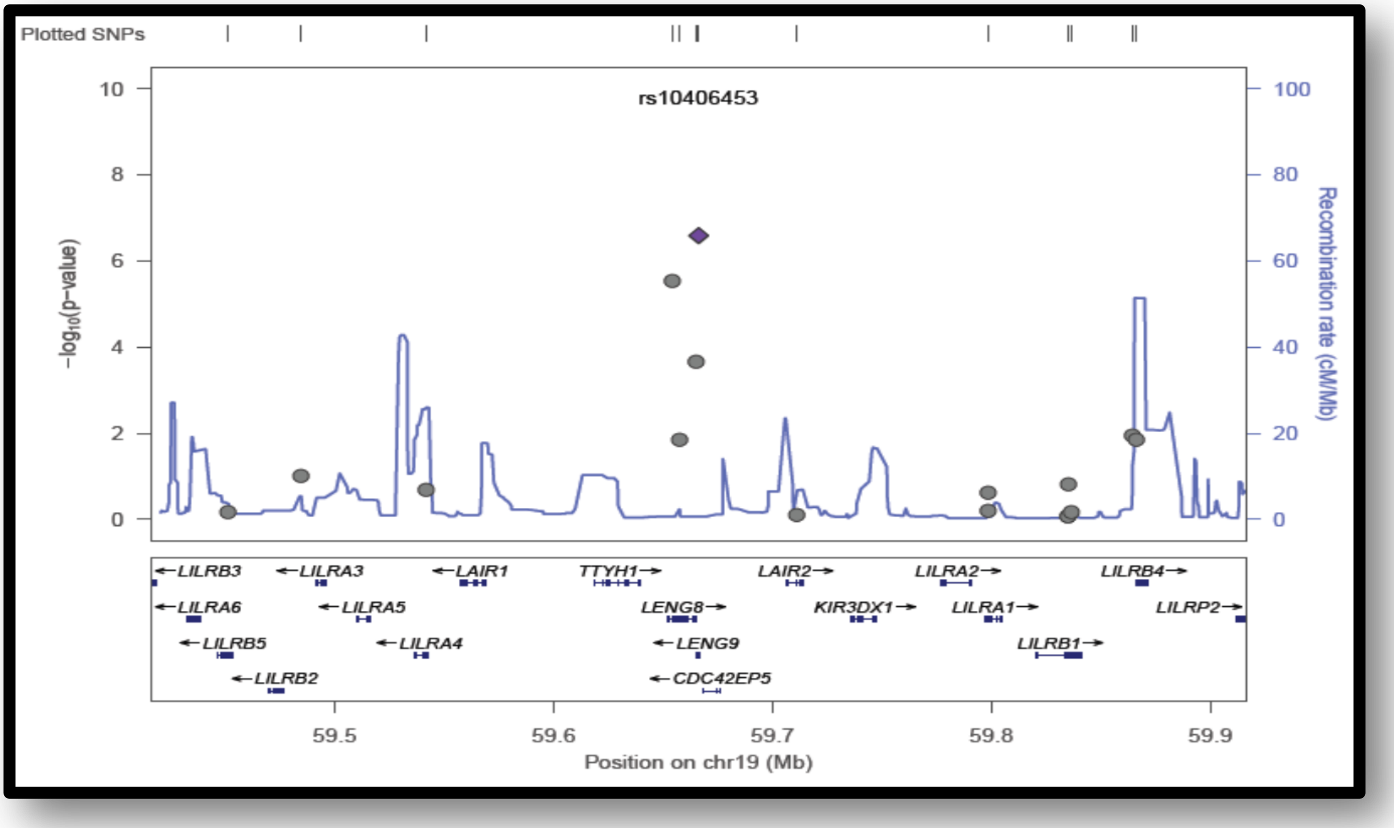


Figure 3-3c. LocusZoom visualization of the region proximal to rs10406453 located in LENG8



<u>Note(Figures 3-3a-c):</u> Linkage disequilibrium key is only shown for regions in which there are genotyped variants in at least marginal LD with the associated variant.

OLS assessment of Significant Median Regression Associations

An important assumption of standard linear regression is that the dependent variable is normally distributed. Due to the characteristically skewed nature of PAI-1 distributions evident in several studied populations, PAI-1 is generally log-transformed in an attempt to bring the distribution closer to normality[31]. To explicitly test whether there was an impact on model inference using log-transformed PAI-1 as the dependent variable in OLS, we tested the five SNPs discovered to be significantly or marginally associated ($p < 10^{-5}$) with PAI-1 through median regression using standard linear regression models adjusted for age, gender, BMI, triglycerides, and PAI-1 4G/5G variant genotype (Table 3-3). For each SNP the reported effects trended in the same direction as that shown in median regression; however in every model, the standard error reported by OLS was greater than that reported by median regression. This resulted in larger 95% confidence intervals and larger p-values reported for each SNP, indicating the increased sensitivity of median regression in a skewed data set (Table 3-3). As further proof of principal, we also decided to evaluate the genetic impact of the PAI1 4G5G promoter polymorphism (rs1799768) which has been shown in multiple populations to have a strong impact on PAI-1 levels (REF). The effect of rs1799768 is linear in nature, and as expected, association results from regression models using linear (rs1799768 effect size: 0.25; p-value: 1.30 x 10^{-4}) and median (rs1799768 effect size: 0.27; p-value: 1.84 x 10^{-4}) were highly concordant.

Table 3-3. Corresponding OLS Results for SNPs marginally/significantly associated with Median PAI-1 levels.

Chr.	Gene	SNP	Beta[1]	SE[2]	95% Confidence Interval		P-value
					LL	UL	
4	*SLC7A11*	rs4479754	0.233	0.062	-0.356	-0.111	2.28E-04
5	*ARSB*	rs1071598	-0.429	0.142	-0.708	-0.151	0.003
7	*CPA2*	rs61997065	0.376	0.137	0.107	0.645	0.006
19	*LENG9*	rs10406453	-0.253	0.095	-0.434	-0.065	0.008
19	*LENG8*	rs1035451	-0.326	0.107	-0.535	-0.117	0.002

Results that remained significant after FDR correction are highlighted in **bold**.

[1] Beta coefficient from median regression model represents the effect of the minor allele; model covariates: age, gender, BMI, triglycerides, and PAI1 4G/5G variant genotype.

[2] SE; Standard error; robust standard errors are reported above.

LL= 95% Confidence Interval lower limit; UL= 95% Confidence Interval upper limit

Discussion

Susceptibility to major thrombotic events is increased by unbalanced or impaired fibrinolysis which is heavily impacted by variation in PAI-1 levels. We hypothesized that common variants significantly impact inter-individual variation in PAI-1 levels. Our results identified three novel variants that remained significantly associated with median PAI-1 after adjustment for environmental covariates and correction for multiple testing. We also explicitly tested the effect of assessing log-transformed PAI-1 levels, where transformation did not bring the distribution to normality, in a moderately sized study cohort. We revealed a significant impact on model inference due to the violation of key assumptions of OLS regression and misspecification of the model; if OLS regression had been employed as our primary analytical method as it is in many studies of PAI-1 levels we would not have discovered key associations with SNPs that may alter the structure or function of genes with plausible biological connections to *PAI-1.*

Of the three SNPs that associated with median PAI-1, rs1071598 was the most significant. Located within the fourth exon of *ARSB,* rs1071518 is responsible for a valine to methionine amino acid change at position 376 (V376M) that is classified as probably benign by PolyPhen (algorithm within SNPInfo) in terms of its effect on ARSB protein function. Although there is no strong evidence that this SNP affects ARSB protein function, the V376M substitution may have a potentially significant effect on structural stability. The substitution of a methionine residue for valine is predicted to cause over packing of protein cores as methionine is a much larger residue than valine, and may be of importance for ARSB protein stability. SNPInfo also revealed that rs1071598 is located within two base pairs of a putative exon splice enhancer motif, perhaps affecting the relative frequency of *ARSB* splice variants.

ARSB has been implicated in the mediation of reactive oxidative species (ROS) production and the activation of ROS-mediated inflammation cascades through interaction with Carrageenan (CGN). CGN is known to induce inflammatory responses in mammals and has been used in model organisms to assess anti-inflammatory therapies[100]. There is also evidence that ARSB has the ability to both replicate and mediate the effects of hypoxia in human tissue[101]. *PAI-1* was recently identified as a hypoxia inducible gene, and has long been established as an inflammatory biomarker[102]. The mutual relationship with inflammatory response between *ARSB* and *PAI-1* presents a potential connection between genetic variants located in *ARSB* and *PAI-1* levels.

The *CPA2* variant, rs61997065, the only significant association with increased median PAI-1, causes a valine to isoleucine substitution at position 67 (V67I). This SNP is proximal to a predicted exon splice enhancer motif, indicating a possible biological role in exon splicing. *CPA2* is a digestive exopeptidase found primarily in the pancreas with evidence of expression in extrapancreatic tissues, including the brain, in both humans and rats[103,104]. Previous studies revealed a possible regulatory role of extrapancreatic *CPA2* in the renin-angiotensin system (RAS) via differential processing of Angiotensin I [105,106]. There are multiple sources of evidence for functional links between the RAS and the fibrinolytic system in humans[107-112]. Additionally, genetic variants of the RAS have been previously discovered to be associated with mean PAI-1 levels in Caucasian and African populations [28,29,55,57,79].

Finally, rs61997065 located in the only exon of *LENG9,* causes an amino acid substitution from histidine to arginine at position 153 (H153R) that is predicted to have benign effect on protein function. *LENG9* encodes a protein that is a part of the leukocyte receptor complex (LRC), an extended gene region on chromosome 19 that is comprised of a large set of genes that encode immunoglobulin superfamily receptors [113]. Although *LENG9* has been mapped to the LRC, its function has not been explicitly investigated, and the relationship between variation at this locus and *PAI-1* remains unclear, making the interpretation of this association difficult.

Although clearly impacted by genetic variation, *PAI-1* levels have a complex genetic architecture that can be revealed using a variety of analytical tools and by comparing diverse populations. Our study identified novel variants associated with *PAI-1* levels overall, and to our knowledge is the only study to interrogate a large number of SNP effects on *PAI-1* levels in an African population. We also explicitly show that this approach is better able to identify significant associations . Standard regression may be appropriate for studies with extremely large sample sizes, such as those seen in meta-analyses, when some model assumptions are violated. For individual studies with modest sample sizes such as ours, and a number of other previous studies of *PAI-1* levels in various populations, the impact of performing "standard" analyses when they are not the most appropriate for that specific study can be significant. If we had only employed OLS regression in our study, then the novel variants that we discovered, several of which had predicted functional impact in genes with plausible biological ties to *PAI-1*, would not have passed multiple testing correction. With the heavy emphasis on controlling Type I error (false positives) by using p-values for model inference, it is highly likely that these effects would not only have been missed in this study but also may have been excluded from future studies[114].

B. Genetic Impact of common SNPs on the Upper Quartile of the PAI-1 Distribution

Introduction

While median regression is superior to OLS regression in the context of our study population, as demonstrated by the studies performed in part A of this chapter, both analyses assume that the relationship between SNP genotype and PAI-1 variation is uniform across the phenotypic distribution. There is no theoretical justification or prior knowledge to support this assumption, and recent evidence from genetic studies of other biological traits highlight the possibility that genetic variants exhibit quantile-specific effects [115-117]. The idea of variable effects of genetic variants on different subsets of the phenotypic distribution has begun to be evaluated in genetics studies over the past few years. One such study design employed by geneticists involves sampling from the extremes or "tails" of the phenotypic distribution to maximize the probability of finding associations. This study design ignores any possibility of non-uniform SNP effects across the phenotypic distribution, but is intended to serve two other purposes: 1) it increases the power of statistical tests to detect associating variants by amplifying any possible genetic signals, and 2) it allows for the identification of variants that have a significant effect on individuals with the most severe affliction. This latter point may not, however, generalize to the impact of genetic variation on the phenotype as a whole.

Quantile regression provides direct estimation of genetic effects on specified regions of the phenotypic distribution, and when targeted to the extremes of the phenotypic distribution, operates under the same premise of extreme sampling[85]. The utility and versatility of quantile regression is especially highlighted in instances where extremes of a phenotypic distribution are correlated with a particular clinical endpoint, such as in the case of plasma PAI-1 levels. Elevated PAI-1 levels are associated with CVD susceptibility and risk of adverse events[4,46,65]; therefore, evaluation of the relationship between SNP genotypes and PAI-1 levels that is not limited to the average of the phenotypic distribution may provide novel clinically relevant information.

To provide comprehensive evaluation of the impact of genetic variants on a clinically relevant portion of the PAI-1 distribution, we performed quantile regression targeting the upper quartile of PAI-1 measurements. We then compared our results to those outlined in part A of this chapter to determine if significantly associating effects on the upper quartile were quantile-specific, and used bioinformatic data mining techniques to predict the functional impact of discovered associations.

Materials and Methods

Subjects

A detailed description of subject recruitment, and ascertainment of demographic data and biomarker measurements can be found in Chapter III, part A.

Genotyping Scheme

Study cohort genotyping protocols are described in detail in Chapter III, part A.

Quality Control Procedures

Quality control procedures pertaining to preliminary processing of PAI-1 levels, assessment of DNA quality and genotyping performance are explicitly outlined in Chapter III, part A. Briefly, quality control thresholds for study inclusion were genotyping efficiency ≥ 95% (subjects and SNPs), MAF ≥ 0.05. Quality control exclusion criterion was a HWE $p < 0.001$. MAF and HWE p-values of SNPs significantly associated with the upper quartile of PAI-1 are presented in Appendix Table 3.

Statistical analysis

Preliminary Analyses

Preliminary analyses aimed at assessing differences in demographic and clinical factors between the genders, as well as the nature of the PAI-1 distribution both before and after log-transformation are presented in Chapter III, part A. As in the analyses performed in Chapter III, part A, log-transformed PAI-1 levels will be used as the dependent variable in this study.

Upper Quartile Regression

Quantile regression has become increasingly popular in human genetics studies as it provides a non-parametric method, invariant to logarithmic transformations of data, which can incorporate covariate adjustment into analyses of the effects of genetic variants. Recent applications of quantile regression at the extremes of the phenotypic distribution in other complex traits suggest that the impact of genetic variants on phenotype may be different, and in some cases stronger, depending on the quantile of the distribution being assessed[85,118]. Because the upper extremes of the PAI-1 distribution have been shown in previous studies to associate with clinical outcomes, we performed upper quartile regression analyses to assess the impact of single variants within this target region of the PAI-1 distribution.

Quantile regression analyses, with robust standard errors, were performed using the quantreg package in R on the upper quartile of the PAI-1 distribution[92]. Regression models were adjusted for age, gender, BMI, triglycerides, and PAI-1 4G/5G variant genotype. Results were visualized via Manhattan plots created using the qqplots package in R[92]. SNPs were coded as described above (Chapter III, part A, Statistical analysis sub-section). Genotypic distribution of SNPs found to be significantly associated with the upper quartile are presented in Appendix Table 4. An FDR threshold of $p < 2.57 \times 10^{-6}$ was calculated as described in Chapter III, part A and used to determine statistical significance.

For gene regions that were found to contain more than one associating variant that passed multiple testing correction, pairwise LD was assessed between associating markers in this region using Haploview[119].

Bioinformatic / Data Mining Investigation of Associating markers

Significant variants were evaluated for putative functional consequences using the web-based SNPInfo database[95]. The specific algorithms within the program that were utilized in this study are fully described in Chapter III, part A.

Results

Quantile regression analyses were performed in the upper quartile (75th percentile) of the PAI-1 distribution. Analyses were adjusted for gender, age, BMI, triglycerides, and 4G/5G variant genotype. Nineteen variants were significant after correction for multiple testing (Table 3-4, Figure 3-3).

The most significant effect in the upper quartile of the PAI-1 distribution was due to rs4755779 located on chromosome 11 ($p= 1.44 \times 10^{-10}$), while the strongest positive and negative effects on PAI-1 were due to rs10462021 (β = -0.434) located on chromosome 1, and rs116307792 (β =0.249) located on chromosome 3, respectively (Table 3-4). Of note, a 72.6kb region on chromosome 11 containing both the pleckstrin homology-like domain, family B, member 1 (*PHLDB1*) and trehalase (*TREH*) genes (*PHLDB1/TREH* gene region) harbored three SNPS that were significantly associated with PAI-1 in the upper quartile; two of which, rs7389 and rs519982, remained significant after correction for multiple testing.

Table 3-4. Upper Quartile Regression Results for Single Variant association with PAI-1 levels

Chr.	Gene	SNP	Beta[1]	SE[2]	95% Confidence Interval		P-value
					LL	UL	
1	*COL16A1*	rs72887331	-0.277	0.053	-0.381	-0.172	**2.64E-07***
1	*FHAD1*	rs12126178	-0.133	0.026	-0.184	-0.083	**2.64E-07***
1	*PER3*	rs10462021	-0.434	0.091	-0.613	-0.256	**2.07E-06**
2	*PLECKHB2*	rs6713972	-0.234	0.037	-0.308	-0.161	**5.14E-10***
3	--	rs13314993	0.212	0.041	0.132	0.293	**2.85E-07***
3	--	rs33483	0.202	0.045	0.113	0.291	9.48E-06
3	*SLC15A2*	rs116307792	0.249	0.047	0.158	0.340	**1.08E-07***
5	*ADAMTS12*	rs61757473	-0.375	0.063	-0.500	-0.252	**3.26E-09***
5	*RAPGEF6*	rs61757473	-0.178	0.039	-0.254	-0.103	4.43E-06
6	*TAGAP*	rs35263580	-0.198	0.033	-0.263	-0.133	**4.14E-09***
7	--	rs2023783	-0.432	0.079	-0.586	-0.278	**4.97E-08***
9	*DBH*	rs4531	-0.195	0.038	-0.270	-0.120	**4.19E-07***
9	*NMRK1*	rs35472028	-0.274	0.059	-0.391	-0.158	4.19E-06
11	*EXT2*	rs4755779	-0.213	0.033	-0.277	-0.148	**1.44E-10***
11	*PHLDB1 / TREH*	rs7389	-0.252	0.049	-0.349	-0.155	**3.70E-07***
	PHLDB1 / TREH	rs2077173	-0.256	0.055	-0.363	-0.149	3.05E-06
	TREH	rs519982	-0.259	0.051	-0.360	-0.158	**5.75E-07***
12	*OR1OP1*	rs76940436	-0.268	0.046	-0.359	-0.177	**1.00E-08***
12	*P2RX7*	rs34219304	-0.302	0.062	-0.423	-0.181	**1.09E-06***
14	*FAM161B*	rs34834232	-0.258	0.053	-0.361	-0.155	**1.13E-06***
14	*NID2*	rs2273430	-0.239	0.050	-0.338	-0.141	**2.32E-06**
16	*C1QTNF8*	rs73494080	-0.283	0.057	-0.395	-0.170	**9.82E-07***
17	--	rs4796217	-0.211	0.046	-0.301	-0.121	4.94E-06
17	*CEP95*	rs9910506	-0.327	0.066	-0.457	-0.198	**8.70E-07***
17	*SECTM1*	rs113432525	-0.276	0.060	-0.394	-0.157	5.63E-06
19	*ERVV-1*	rs10403404	-0.201	0.045	-0.289	-0.114	7.60E-06
19	*PDE4C*	rs1444689	0.224	0.050	0.125	0.322	9.16E-06
22	*C22orf43*	rs75824255	0.185	0.040	0.106	0.263	4.92E-06

Results that remained significant after FDR correction are highlighted in **bold**; only a subset of significant results are presented above (p-value ≤ 10^{-6})

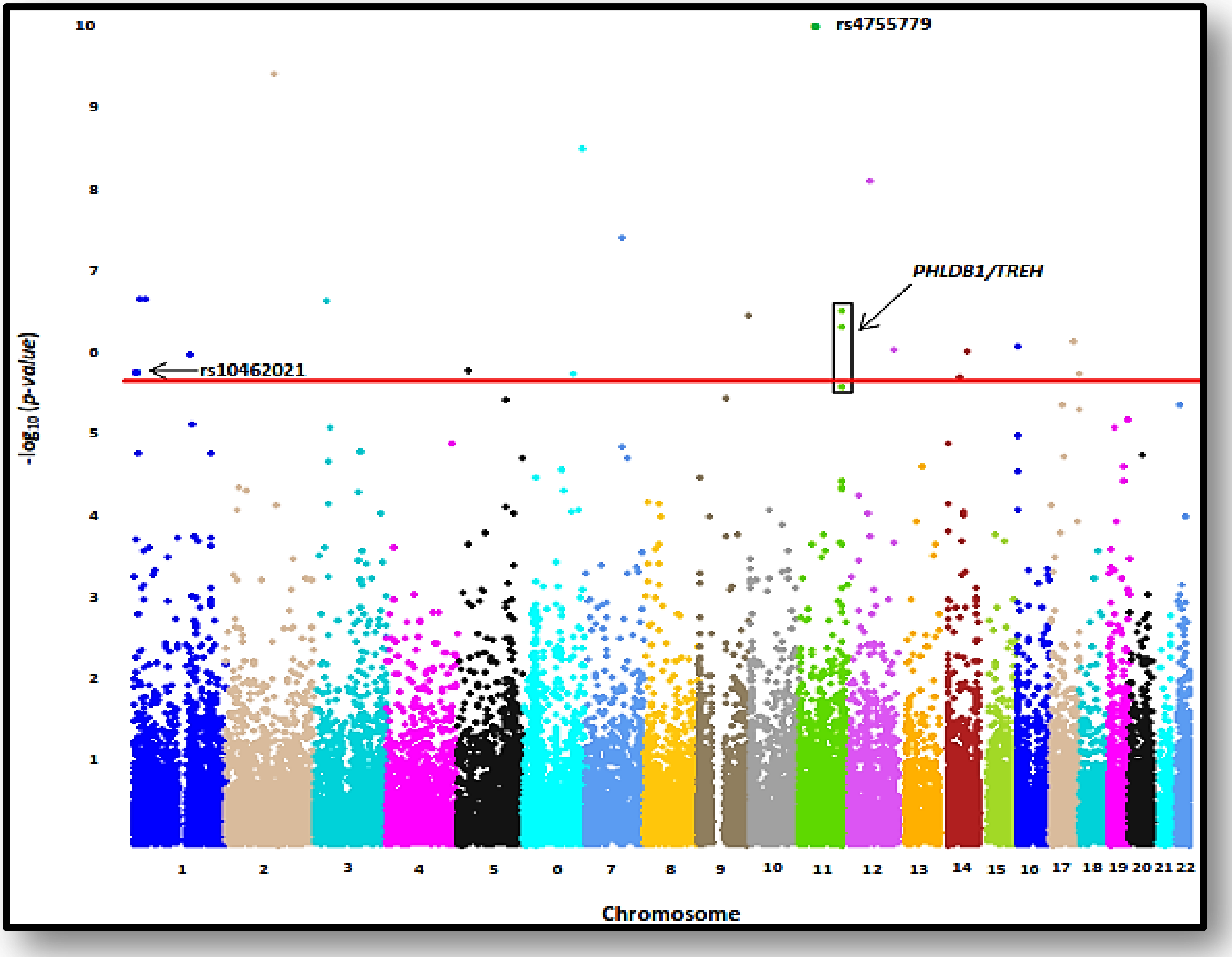
[1] Beta coefficient from median regression model represents the effect of the minor allele; model covariates: age, gender, BMI, triglycerides, and PAI1 4G/5G variant genotype.

[2] SE; Standard error; robust standard errors are reported above.

LL= 95% Confidence Interval lower limit; UL= 95% Confidence Interval upper limit

*Effect remained significant after Bonferroni correction (threshold = 1.29E-06)

Figure 3-4. Manhattan Plot of SNP Association Analysis with Upper Quartile of PAI-1 distribution



Note: Only markers on chromosomes 1-22 are presented above; regions from the X chromosome, Y chromosome, pseudo-autosomal region of the X chromosome, and mitochondrial markers have been excluded. The three most significant markers are labeled in bold. Red Line represents FDR significance threshold (2.57 x 10^{-6}). Loci with more than one significant variant are signified by a box. Noteworthy significant associations are labeled in bold.

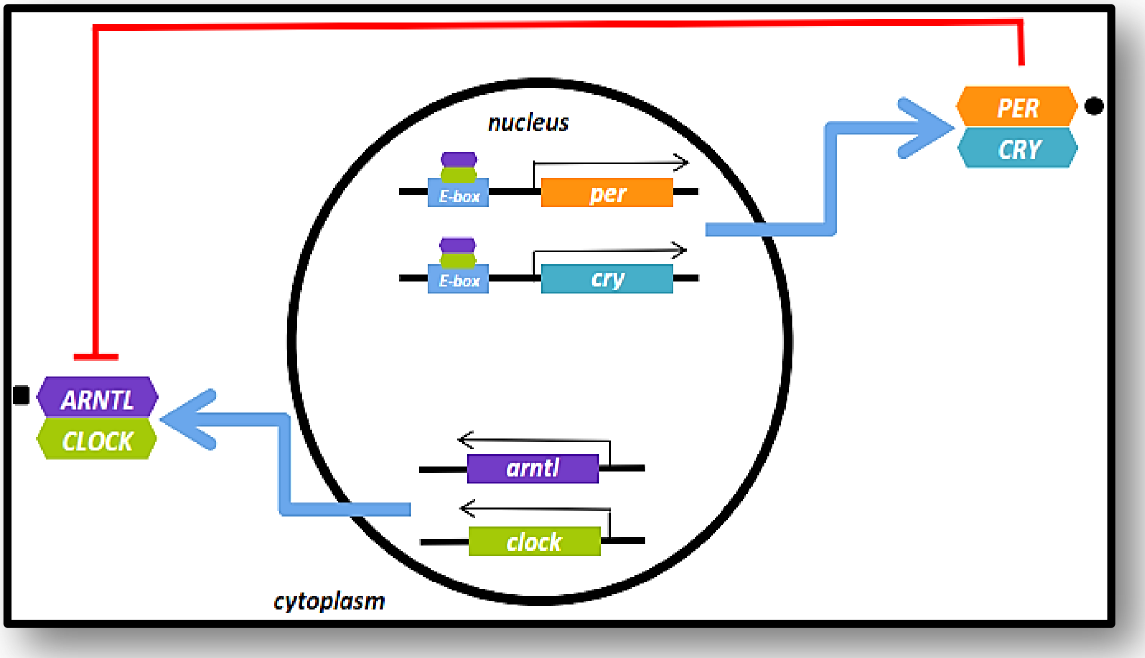
Discussion

Upper quartile analyses revealed 19 associating variants; of particular note among these associating markers were 1.) two non-synonymous SNPs located in genes with a credible connections to PAI-1, rs4755779 in *EXT2* and rs10462021 in *PER3*, respectively, and 2.) three SNPs located in the *PHLBD1/TREH* gene region on chromosome 11.

The *EXT2* SNP, rs4755779, is a missense variant that causes a methionine to valine substitution at position 42 (M42V) predicted to have a benign effect on protein function by SNPinfo. *EXT2* encodes a protein involved in heparin sulfate biosynthesis, and has been associated with hereditary multiple exostoses and Type 2 diabetes[120-122]. There is a plausible biological connection between *EXT2* and *PAI-1* via heparin-binding growth factors (HBGF); HBGFs have been implicated in the modulation of *PAI-1* expression. In particular, HBGF-1 has been shown to inhibit *PAI-1* expression in human umbilical vein endothelial cells[123].

Rs10462021, a missense variant in *PER3* responsible for a histidine to arginine substitution at position 1149 (H1139R), is predicted to have an effect on protein function although the nature and mechanism of this effect remains unclear. *PER3* is a member of the circadian rhythm pathway and has been shown to affect inflammatory response through increasing the secretion of inflammatory cytokines[124]. Previous studies in model organisms also report an association between *PER3* and susceptibility to CVD; specifically, transgenic *PER3* knockout mice show increased susceptibility to arteriosclerotic disease[125]. The identification of rs10462021 in *PER3* is particularly noteworthy due to the recent discovery of variants in another prominent member of the circadian rhythm pathway*, aryl hydrocarbon receptor nuclear translocator-like gene* (*ARNTL*), that are significantly associated with plasma *PAI-1* levels in a recent meta-analysis performed in Caucasians [31]. PER3 and ARNTL are major regulators of the circadian clock mechanism, a transcriptional timing apparatus governed by multiple positive and negative feedback loops[126] (Figure 3-4). The interaction between the PER3/CRY and ARTNL/CLOCK heterodimers is particularly noteworthy due to the substantial evidence of activation of the *PAI-1* promoter by ARNTL/CLOCK and its direct impact on *PAI-1* expression[81,127].

Figure 3-5. Schematic of the Circadian Rhythm Pathway.



Note: Figure above was adapted from Stow *et al.* 2011[128] .

- ■ Significantly associated with mean PAI-1 levels in Caucasian populations
- ● Significantly associated with PAI-1 levels in the current study of Ghanaians

The discovery of significantly associated variants in two separate genes, *ARNTL* and *PER3*, belonging to the same biological pathway in both African and Caucasian populations highlights the possibility of population-specific vs. universal gene effects. The individual gene effects of *PER3* and *ARNTL* on PAI-1 variation may be population specific, but the involvement of the circadian rhythm pathway may be generalizable. A difference in allele and genotype frequencies at the *PER3* variant, rs10462021, may be responsible, in part, for a population-specific effect on *PAI-1* levels at this locus. A previous study by Ciarleglio *et al.* investigating patterns of genetic variation in human circadian rhythm genes revealed a significant difference in allelic and genotypic distributions of rs10462021 between Ghanaian (some of which overlapped with our cohort) and European American (EA) participants; the A allele was more prevalent in the Ghanaian population (freq. = 0.96) than in the European American population (freq. = 0.83) [129]. There was no significant difference between allelic or genotypic frequencies for this variant between Ghanaian and African American participants[129]. These results indicate that, although there may be population-specific gene effects, these effects may function through a common physiological pathway.

In addition to uncovering associations with variants in several novel genes, a 72.6kb region on chromosome 11, containing two genes *PHLBDI* and *TREH,* contained multiple variants associated with the upper quartile of *PAI-1.* Of the three variants identified in the *PHLDB1/TREH* gene region, two SNPs, rs519982 and rs7389, passed correction for multiple testing. These three SNPS are all in high linkage disequilibrium (LD) with each other ($0.94 < r^2 < 0.97$), indicating that these variants represent a single association signal. Therefore, functional predictions are virtually impossible. However, we can speculate based on the putative individual SNP functions. Rs519982 is located in a region predicted to contain a transcription factor binding motif 14.9kb upstream of the *TREH* start codon. Its predicted location in a transcription factor binding site proximal to the *TREH* gene boundary may have functional implications as variation at comparable loci have the potential to effect gene expression. The second SNP in the *PHLDB1/TREH* gene region, rs7389, is a located in the 3' UTR of *PHLDB1* that is predicted to affect microRNA (miRNA) binding site activity; miRNAs are single stranded RNA molecules that can inhibit protein translation[130]. Although there is little evidence of a direct connection between *TREH/PHLBD1* and *PAI-1*, the three genes do share some overlapping characteristics, as well as disease associations. Similar to *PAI-1*, *TREH* has been identified as a stress response gene and has been shown to associate with susceptibility to Type 2 diabetes[131-134]. Likewise, both *PHLDB1* and *PAI-1* have been implicated in the etiology of glioma[135,136].

Elevated *PAI-1* levels, in particular, are associated with increased susceptibility to CVD and in some cases severity of disease[2,4,137]. Performing upper quartile analyses allowed an explicit test of the association between single variants and elevated PAI-1 levels and revealed novel associations which did not overlap with the significant effects on median PAI-1 presented in Chapter III, part A, demonstrating the quantile-specific genotype-phenotype relationships between SNPs and variation in PAI-1. Performing upper quartile analyses provided us with additional insight into a clinically relevant subset of the *PAI-1* distribution, which has been understudied, especially in West African populations. Increasing our understanding of the impact of genetic variation on PAI-1 levels at the higher end of the distribution may aid in the development of targeted therapies that may not be effective in the general population, but will have a significant impact on a subset of the population already at increased risk of CVD. Additionally, the discovery of *PER3*'s association with upper quartile *PAI-1* levels provides further evidence not only of the involvement of circadian pathway members in *PAI-1* regulation, but also indicates that the effects of the circadian pathway may be generalizable between Caucasian and West African populations, even if the specific gene effects differ.

CHAPTER IV

PRESENTATION OF A NOVEL APPROACH TO IDENTIFY CANDIDATE GENES BASED ON MULTI-LOCUS ASSOCIATION SIGNALS

Overview

Identification of candidate genes likely to harbor functional variants that impact phenotype is a logical next step in unraveling the convoluted interplay between genetic variants that characterize complex disease. We present a novel approach that incorporates standard gene prioritization methods from GWAS, candidate gene, and QTL mapping study designs that can be applied to SNP level data to aid in the selection of candidate genes likely to harbor functional variants with biological connections with a given phenotype. We evaluated our results to find pairwise models that were identified as the "best" model for multiple gene regions; the summation of these gene regions defines our new loci of interest, and analogous to QTL mapping, we then identified all genes within this region as candidate genes for PAI-1. In addition to identifying novel candidate genes to aid us in prioritizing future studies, we also identified four statistically significant synergistic interaction effects ($p < 0.001$) associated with PAI-1 levels. We performed a preliminary evaluation of the novel candidate genes for PAI-1 variation that we identified using our novel approach and present possible biological / mechanistic connections. A review of current literature revealed that all four of our identified regions of interest not only contained genes with intuitive biological connections to PAI-1, but each region was found to lie within at least one previously discovered QTL region for cardiovascular disease related traits such as early onset myocardial infarction or biomarkers related to variation in PAI-1 such as triglycerides or body weight. We go on to further interrogate one of our identified loci of interest (located on chromosome 5) to demonstrate the intuitive connections to be found between genes within this region of interest and PAI-1. We also characterized the four statistically significant pairwise effects identified by qMDR that were used to define our regions of interest.

A. Presentation of Multi-lOcus based selection of Candidate Genes (MOCA)

Introduction

One of the major goals of human genetic studies is to discover significant associations between genetic markers and complex disease/traits. Arguably, the most challenging problem facing researchers in this endeavor is elucidating the connections between discovered variants, their biological mechanisms, and their impact on phenotype. One method of incorporating information gained from single variant analyses in a way that is more biologically translatable is to map them to the nearest gene(s) and then consider the gene as a candidate for further evaluation in the context of a given phenotype. This method recognizes that the identified SNP is not likely to be the causal variant but is in LD with the true variant, and that this variant(s) may lie within or near the same gene as the identified marker. Examination of the entire gene region may reveal other SNP-level associations and perhaps uncover a biological connection with variation in phenotype at the gene level. Three widely practiced methods used to identify candidate genes in complex disease are 1.) genome wide association studies 2.) candidate gene studies (genes chosen based on biological knowledge) and 3.) genome-wide linkage analyses (quantitative trait loci (QTL mapping)); all of which have certain weaknesses when applied to complex disease.

Genome-wide association studies (GWAS) perform large numbers of agnostic single variant tests (SNP-level) to determine association between these variants and a given phenotype. Significant variants are then mapped by their chromosomal location to the nearest gene; this gene is then considered to be a putative candidate gene for the phenotype of interest and further studies may be performed within the gene region in an attempt to discover biological links between the new candidate gene and phenotype. This method suffers from several weaknesses; 1.) the SNP may not lie within or near a gene region, making gene assignment problematic, 2.) the SNP may lie within a gene region but the observed effect with phenotype is due to its influence on another gene outside of the region in which it is physically located, 3.) unless multiple transcripts overlap, each significant SNP association will only yield one new candidate gene to be interrogated and perhaps most significantly, 4.) recent evidence in multiple phenotypes has shown that epistasis (SNP-SNP interactions) may play a more important role in variation in complex phenotypes than single variants[138,139].

Candidate gene studies use prior biological knowledge to select a small subset of genes with proven or plausible mechanism regarding a given phenotype. Variants within the selected genes are then evaluated for association with phenotype, and those found to have significant associations undergo further investigation into possible functional effects at the gene level, as the connection between gene and phenotype was already established *a priori*. The major weakness of this approach is that it does not allow for the discovery of novel candidate genes; genes are chosen for study because they already showed evidence of connection with the phenotype.

Another classic method of identifying candidate genes that may harbor biologically functional variants associated with phenotype is linkage analysis, which in recent years has been extended to genome-wide linkage analysis. Linkage analysis has historically been used to identify quantitative trait loci predicted to harbor genes associated with a given phenotype within families. Linkage analyses have been largely successful when applied to monogenic (Mendelian) disorders, and in recent years has been applied to complex traits such as CVD. Unfortunately, when applied to more complex phenotypes, interpretation of QTL analyses becomes daunting as identified regions tend to be expansive, harboring as many as 100-200 genes[140,141]. And as with candidate gene and GWAS studies, QTL for complex traits discovered via linkage analyses also suffer from lack of replication. A recent example of this is the variable results obtained from genome-wide linkage analysis studies performed to identify QTL in atherosclerotic disease regulating traits[142]. Over 40 QTLs have been identified for atherosclerosis, but only three of these have replicated; additionally, it is worth noting that most of the single variant associations identified through GWAS did not fall within any previously identified atherosclerosis QTL[142].

CVD is a complex disease due, in part, to impaired or unbalanced fibrinolysis, the process by which clots are degraded, which is heavily regulated by PAI-1[2,4]. Genetic impact on PAI-1 levels has been primarily studied using methods that assess the marginal effects of single SNPs in several populations; however, the loci discovered using this study design explain only a small portion of the inter-individual variation. A first step beyond single SNP analyses is to identify regions of interest more precisely and prioritize them for more intensive study. We present a novel approach, Multi-lOcus based selection of CAndidate genes (MOCA) that incorporates the unconventional usage of the quantitative Multifactor Dimensionality Reduction (qMDR) algorithm as a part of an pipeline to identify regions of interest harboring putative candidate genes to be prioritized for inclusion in future studies. MOCA has the advantage over standard single variant based methods in that it uses multiple SNPs within a region to effectively localize regions of interest.

Materials and Methods

Subjects

A detailed description of subject recruitment, and ascertainment of demographic data and biomarker measurements can be found in Chapter III, part A.

Genotyping Scheme

Study cohort genotyping protocols are described in detail in Chapter III, part A.

Quality Control Procedures

Quality control procedures pertaining to preliminary processing of PAI-1 levels, assessment of DNA quality, genotyping efficiency (subjects and SNPs), HWE and MAF thresholds are outlined in detail in Chapter III, part A. In addition to these procedures, further processing of genotyped variants was required in preparation for downstream interaction analyses. The presence of LD between markers decreases the power of qMDR to detect significant interaction effects. In order to optimize qMDR performance and decrease redundancy in downstream statistical models, SNPs in high LD were filtered using the LD prune feature in PLINK[88] according to the following protocol; for each chromosome a window of 50 SNPs was evaluated for pairwise LD using the correlation coefficient (r^2), for every pair of SNPs determined to be in high LD ($r^2 > 0.8$) one SNP was removed, and the window was shifted by 5 SNPs along the chromosome and the procedure was repeated. After LD based SNP pruning, 34,418 SNPs remained for inclusion in interaction analyses.

Description of MOCA

Rationale

With the variable success of GWAS studies to explain more than a small portion of the variation in complex disease, recent years have seen a paradigm shift away from the notion that an as yet undiscovered subset of single common variants accounted for the majority of the genetic impact on complex diseases towards ideas that include the hypothesis that a greater portion of the heritability in complex traits is attributable to non-linear (epistatic) interactions between genetic variants. This has led to the development of numerous analytical techniques aimed at detecting and characterizing the effects of both linear and non-linear interaction effects with and without the presence of significant marginal effects to provide a more comprehensive understanding of the genetic architecture underlying complex traits[138,139].

Single locus analysis using median regression (Chapter III, part A) identified three significant SNP-level effects. We then followed the standard procedure of mapping these variants to their associated genes by chromosomal position and attempted to draw connections between the aforementioned genes and PAI-1. Although these analyses revealed novel associations, in the context of gene discovery, we interrogated over 38,871 single variants and only identified three novel candidate genes for inclusion in future studies. Considering the possible importance of multi-locus effects (interaction effects) in contributing a significant portion of the heritability in complex disease, the standard associated SNP – to – gene approach in prioritizing candidate genes based on the results of association studies may not be the most inclusive approach. Single SNP analyses may not adequately capture associations in genic regions because the ability of a single SNP to indirectly capture functional variants is not as powerful as multiple SNP approaches; this would be analogous to imputation using a single SNP vs imputation using a haplotype. Hence, multi-locus associations within a genic region may be more effective at tagging functional variants within that region. Given that we genotyped our samples using a gene based platform (HumanExome Chip) as described in detail in Chapter III, we have the ability to assess multiple variants within or proximal to genic regions throughout the genome. This approach, because it is gene based, also provides improved ability to define putative mechanisms.

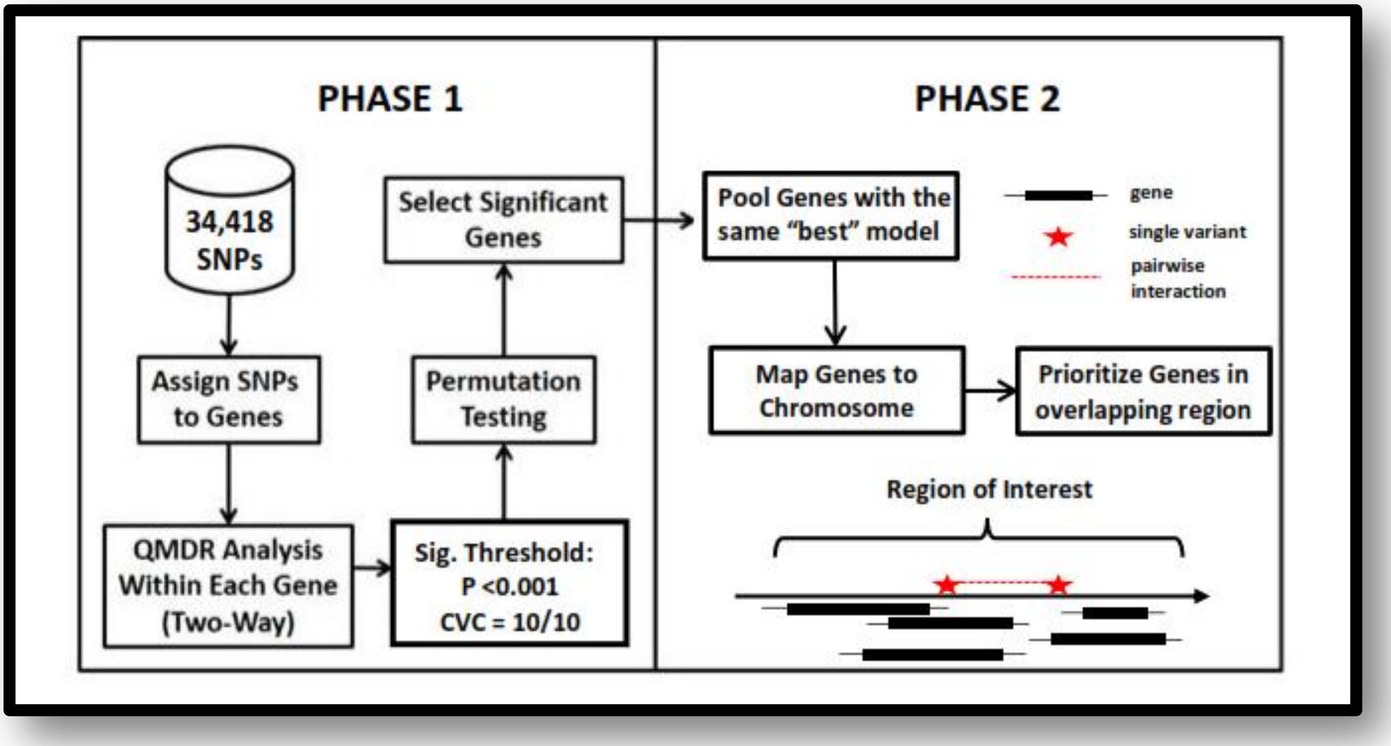
General Summary of MOCA

We propose a novel approach, Multi-lOcus based CAndidate gene finder (MOCA), that incorporates quantitative Multifactor Dimensionality Reduction (qMDR) analyses as a part of a pipeline to identify and prioritize candidate genes obtained from multi-locus analyses in SNP-level data. QMDR is a powerful non-parametric method that can detect multi-locus effects, both additive and non-additive, in the absence of single site main effects. The traditional application of qMDR to association studies is to identify the most informative interaction effect overall in a set of genotyped markers. By first mapping our SNPs to genic regions, and performing qMDR analysis within each region independently, we will create an environment in which qMDR is identifying the most informative genic model for overlapping subsets of the genome defined by gene regions. This will allow us to detect multiple regions harboring associating variants. When these independent results are then interpreted together, they will identify a larger region which may harbor additional gene regions of interest not discovered during qMDR analysis.

To incorporate the intuitive connections between loci and phenotypes observed in candidate gene studies, we first mapped our 34,418 common SNPs (described in Chapter III, part A) to genes via their chromosomal location. We further extended each formal gene boundary by 500kb at the 5' and 3' end to capture putative regulatory regions, and then repeated our SNP to gene mapping procedure. To avoid the gene selection bias seen in candidate gene studies, we allowed SNPs to be mapped to multiple genes. To incorporate the agnostic SNP association characteristic of GWAS analyses, while acknowledging the likelihood that multi-locus (SNP-SNP) effects may be more informative in the context of complex disease association than single variants, we performed independent pairwise analyses within each gene region using qMDR to identify the most significant model for each gene region. Permutation testing was performed in each gene region separately and a significance threshold of cross validation consistency (CVC) = 10/10 and permutation p < 0.001 was used to judge significance of pairwise qMDR models. Genes regions in which the same pairwise model was significantly associated with PAI-1 levels were pooled to form one large overlapping region, and all genes falling within this region were considered to be putative PAI-1 candidate genes and were equally prioritized for inclusion in future analyses. Genes located within loci of interest were then preliminarily evaluated for plausible biological connections to PAI-1.

Although the main aim of MOCA is to aid in the identification and prioritization of candidate genes from SNP-level association studies, an additional benefit of the approach is the possible discovery of strong truly intragenic epistatic (synergistic) effects on the phenotype of interest. These results can then be evaluated and potentially used to direct higher-order (three/four locus effects) or pathway level approaches. A schematic of the MOCA pipeline is presented below:

Figure 4-1. Schematic of MOCA approach.



Statistical analysis

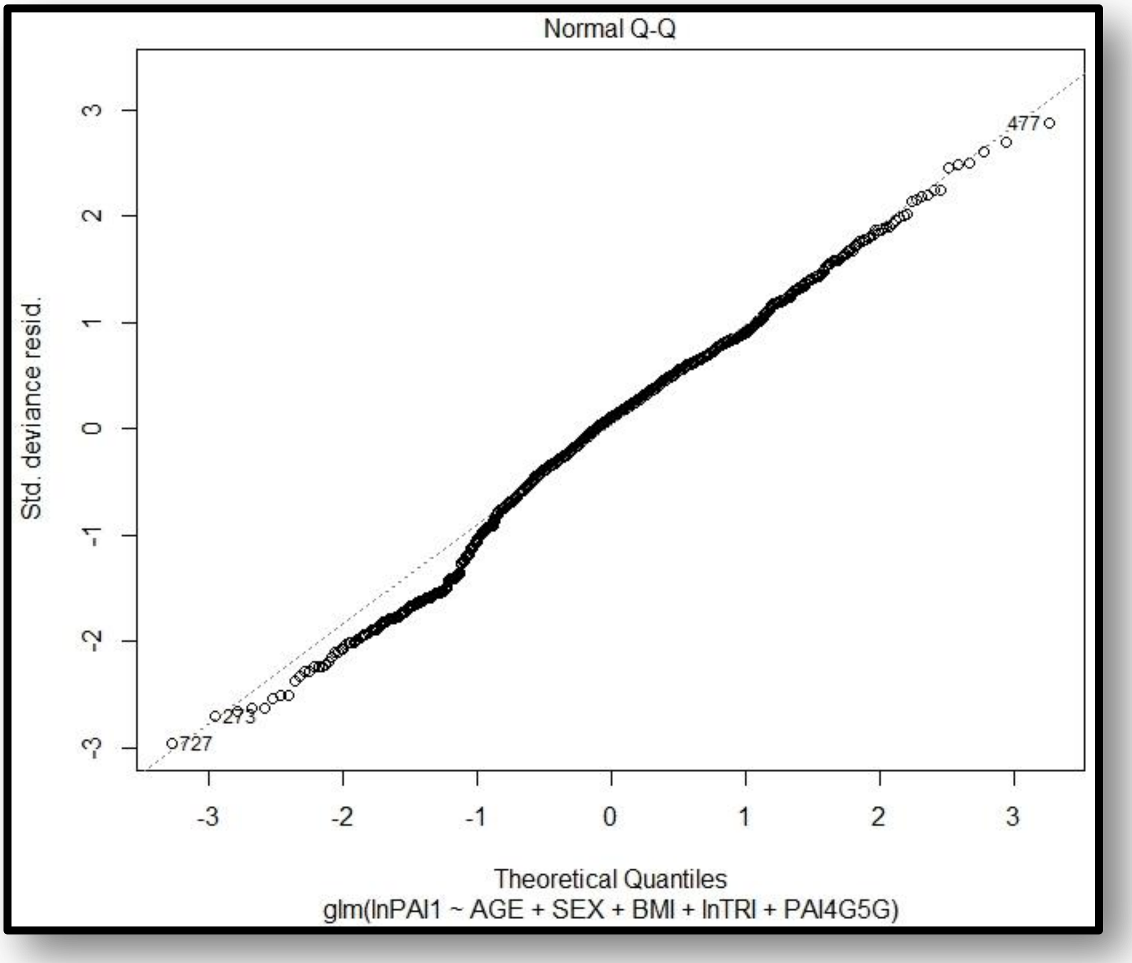
Data Processing

MOCA integrates qMDR analyses as a part of a pipeline to prioritize candidate genes from multi-locus association testing. QMDR is a non-parametric approach that can be applied to continuous traits that is able to detect interaction effects with and without the presence of significant main effects[143]. Although qMDR is optimized for the use of continuous phenotypic data, the algorithm is currently unable to adjust for the effects of covariates in the evaluation of interaction effects[143]. There are several demographic, genetic, and biological factors that are known to strongly impact variation in PAI-1 levels; BMI, age, gender, triglycerides, and genotype at the PAI-1 4G/5G promoter variant [33,34,41,42,44,47,51,55]. In order to appropriately evaluate the effect of epistatic interactions on variation in PAI-1, our model must incorporate and adjust for the aforementioned covariates. To accomplish this, prior to qMDR analysis, linear regression was performed using the following regression model:

lnPAI-1 ~ Age x BMI x lnTRI x Gender x PAI-1 4G/5G genotype

After performing linear regression, model residuals were recorded for use as the dependent variable in downstream qMDR analyses. Using regression model residuals as input for qMDR allows for the assessment of the trait of interest after accounting for the main effects of significant covariates. This procedure has been previously published and has been shown through simulation studies to be an effective method of accounting for the effect of potentially confounding variables in SNP interaction association analyses[143]. Normality of residuals was assessed using the Shapiro Wilkes test ($p > 0.05$) and was visually assessed using a Quantile – Quantile plot, with the normal distribution as the reference (Figure 4-1).

Figure 4-2. Quantile-Quantile Plot of Residuals for use as Dependent Variable in qMDR Analyses.



It is important to note that in the calculation of residual values linear regression only considers observations with complete information available for all included model covariates. Due to the strong impact of the included regression model covariates on variation in PAI-1 levels, individuals who were missing demographic information for any of these factors and therefore unable to provide a residual score were excluded from qMDR analyses. After filtering of excluded individuals, 925 individuals (388 males, 537 females), remained for inclusion in qMDR analyses, all of whom were from the 992 previously studied cohort as only these had direct measures of the 4G/5G genotypes[28,57]. Distributions of demographic and biomarker measurements for these individuals were compared by gender using the STATA 11[90] software program and are presented in Table 4-1, and visualizations of the PAI-1 distribution (ng/ml) before and after log-transformation for these individuals are presented in Figures 4-3a and 4-3b.

Table 4-1. HeART cohort gender-separated demographic and clinical characteristics.

		Males (n=388)	Females (n=537)	P-Value[1]
Age (years)*		43.88 (12.56)	42.66 (10.80)	0.240
Body Mass Index (kg/m^2)*		23.91 (4.24)	26.62 (5.32)	<0.001
Triglycerides (mg/dL)*		97.22 (61.72)	85.26 (44.24)	0.013
serum PAI-1 levels (ng/mL)*		6.74 (8.98)	6.89 (7.62)	0.013
PAI-1 4G/5G genotype	4G/4G	20	34	0.264[2]
	4G/5G	108	171	
	5G/5G	260	332	

*; mean (standard deviation) untransformed variables are presented

[1] P-values are from the Wilcoxon Rank Sum test unless otherwise indicated.

[2] P-values are derived from the Chi-square test of association.

Figure 4-3a. PAI-1 Distribution in Individuals included in SNP-SNP Interaction Analyses.

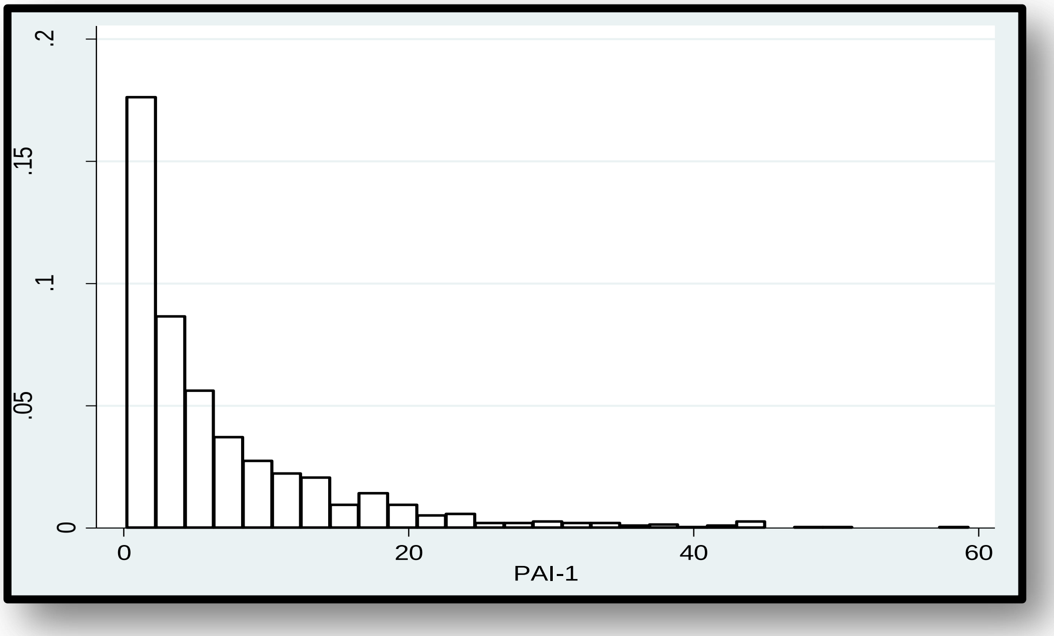
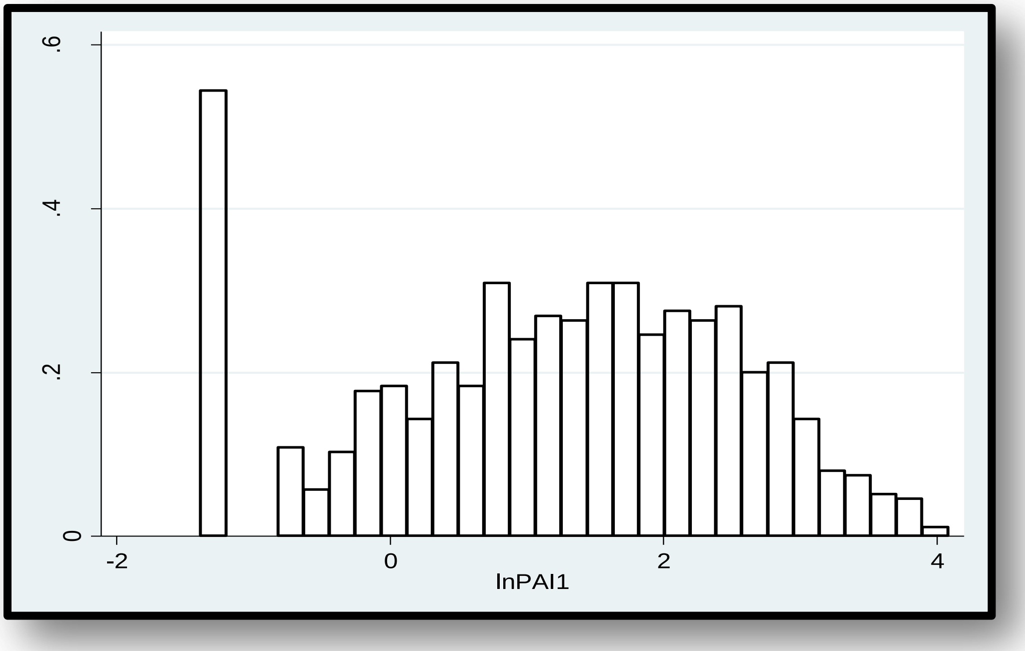


Figure 4-3b. PAI-1 Distribution after log-transformation in individuals included in SNP-SNP Interaction Analyses.



All data processing analyses were performed using the R software package unless otherwise noted[91]. We then recorded the residuals from the OLS regression to be used as the dependent variable in qMDR analyses.

SNP to Gene Region Mapping Protocol and Rationale

As a first step towards identifying loci of interest that are likely to harbor variants associating with PAI-1 levels, we mapped our genotyped variants (n=34,418) to 21,041 gene regions based on chromosomal location; these regions were constructed by extending formal gene boundaries defined by the UCSC genome browser[144] 500kb in the 5' and 3' directions. SNPs were assigned to gene regions using this protocol to target intragenic effects while also increasing the probability of capturing pertinent regulatory elements not located within the formal gene boundaries. An additional benefit of this SNP-to-gene mapping scheme is that it increases the possibility that pairwise effects will tag multiple gene regions as their boundaries are likely to overlap; this will in turn result in the gene region defined SNP sets interrogated individually by qMDR to be highly overlapping as well.

Our mapping strategy allows for the assignment of individual SNPs to multiple gene regions to prevent gene assignment bias and allow for the possibility that regulatory elements affecting one gene may lie within the boundaries of another. One caveat of allowing non-unique SNP assignment to genes is the redundancy that is introduced into the sets of markers assigned to gene regions. If two gene region boundaries are completely or largely overlapping, such that the same subset of SNPs are assigned to both, then redundancy will be introduced into the model as the same subset of variants will be tested multiple times, and multiple gene regions will report identical effects. Due to the aforementioned redundancy, it is highly likely that we tested far fewer independent loci than the 21,041 gene regions that were annotated.

QMDR Protocol and Analysis Pipeline

As previously mentioned, qMDR analysis is an integral part of the MOCA pipeline. QMDR is an extension of the original multifactor dimensionality reduction (MDR) machine learning approach that is optimized to model epistatic interaction effects on quantitative traits[145]. A hallmark of the original MDR algorithm, which is used to evaluate epistatic interaction in dichotomous phenotypes, is the use of constructive induction to identify multi-locus genotypic combinations associated with disease risk and then reduce them into a single new attribute that is used to model the relationship between this newly constructed attribute and case-control status. The major difference between the original MDR approach and qMDR is that the latter compares the mean value of each multi-locus genotype group to the overall mean to determine association with phenotype, while the former evaluates association with phenotype by comparing the case/control ratios of the multi-locus genotypic groups to a fixed threshold[143,145]. Specifically, qMDR incorporates constructive induction into the evaluation of epistatic interactions in the context of quantitative, or continuous, traits as described in Gui *et. al.*[143] and summarized below:

1. *Assume that there are* ***m*** *SNPs in the dataset to be analyzed. To evaluate a* ***J*** *order interaction effect, a subset of* ***J*** *SNPs are selected from the* ***m*** *SNPs in the dataset*

2. *Calculate the mean of the phenotypic trait for each multi-locus genotype combination defined by the* ***J*** *SNPs (genotypic mean) and compare this value to the overall mean calculated from the entire dataset.*

3. *Label the genotypic mean as "high level" if it is greater than the overall mean; label all other genotypic means as "low level". Once all genotype combinations have been labeled as described, a new binary attribute is constructed by pooling genotype groups based on their assigned level.*

After assigning genotype groups to either high or low levels for each **J** SNPs, differences between the mean phenotypic values of the two groups are compared using a Student's T-test and the resulting t-statistic is used as the training score to select the best ***J*** order interaction model. QMDR uses an identical cross validation procedure to select the best overall interaction model and control for model overfitting as that incorporated into traditional MDR; the only difference is that qMDR substitutes the training and testing scores in lieu of the training and testing balanced accuracies utilized in MDR[143].

We performed qMDR analyses to identify the best pairwise SNP-SNP model in each of the annotated 21,041 gene regions, independently. For each gene region, qMDR identified the most informative SNP-SNP model and reported the corresponding testing and training T-statistics, as well as the cross validation consistency (CVC). After qMDR identified the best fit model for each gene region, permutation testing (n=1000) was performed (within the specified gene region) to generate a model p-value. It is important to note that because SNPs were not uniquely mapped to gene regions, the same pairwise model may be the most parsimonious model for multiple gene regions. However, because qMDR assesses model fit by identifying the most informative model among a specific set of SNPs, there may be a difference in the significance reported for the same pairwise model among assigned gene regions. An example of this phenomenon is presented below:

1. ***Assume that the pairwise model SNP_1 – SNP_2 was identified as the best fit model in a region that contained six genes with overlapping boundaries:***

Model	Locus of Interest	Gene	T-statistic Training	T-statistic Testing	CVC	P-value
SNP_1 – SNP_2	Locus 1	A	0.4128	0.4862	8/10	0.026
		B	0.4321	0.5024	9/10	0.003
		C	0.4569	0.5362	10/10	<0.001
		D	0.4569	0.5362	10/10	<0.001
		E	0.4569	0.5362	10/10	<0.001
		F	0.3658	0.2634	4/10	0.087

2. ***QMDR will report identical statistics (T-statistics, CVC, p-value) when SNP sets for different gene regions are identical***

Gene	Mapped SNPs
A	snp1, snp2, snp3,snp4,snp5,**SNP1, SNP2**,snp6
B	snp5,**SNP1, SNP2**,snp6, snp123
C	snp4,snp5,**SNP1, SNP2**,snp6
D	snp4,snp5,**SNP1, SNP2**,snp6
E	snp4,snp5,**SNP1, SNP2**,snp6
F	snp5, **SNP1, SNP2**, snp6, snp123, snp54, snp29, snp07, snp36, snp456, snp678, snp2894, snp0893

The threshold to determine statistical significance of associations between pairwise effects and PAI-1 is a CVC of 10/10 and a permutation p-value ≤ 0.001.

Results

Presentation of Regions Identified using the MOCA Approach

MOCA analysis identified four loci of interest on chromosomes 5,8,17, and 20, which contained a total of 28 novel PAI-1 candidate genes. The pairwise interaction model used to identify these regions as well as the qMDR test statistics for each model is presented in Table 4-2 below.

Table 4-2. Pairwise interaction models significantly associated with PAI-1 levels

Interaction Model	MOCA Regions of Interest[1]	Gene Regions[2]	Testing T- Statistic[3]	Training T- Statistic[4]	P-Value[5]
rs3985058,rs10064163	Chr. 5: 111642442 - 113422334	*REEP5*	4.6463	4.8973	< 0.001
		SRP19	4.6463	4.8973	< 0.001
		APC	4.6463	4.8973	0.001
		EPB41L4A-AS1	4.6463	4.8973	0.001
		SNORA13	4.6463	4.8973	0.001
		EPB41L4A	4.6463	4.8973	0.002
rs925030,rs17054477	Chr. 8: 24684772-26545124	*CDCA2*	3.9571	4.0736	< 0.001
		EBF2	3.9571	4.0736	< 0.001
		GNRH1	3.9571	4.0736	< 0.001
		KCTD9	3.9571	4.0736	< 0.001
		DOCK5	3.9571	4.0736	0.003
rs9907759,rs2270517	Chr. 17: 7704731 - 8822516	*ARHGEF15*	5.1776	5.4569	< 0.001
		AURKB	5.1776	5.4569	< 0.001
		CTC1	5.1776	5.4569	< 0.001
		LINC00324	5.1776	5.4569	< 0.001
		PFAS	5.1776	5.4569	< 0.001
		RANGRF	5.1776	5.4569	< 0.001
		SLC25A35	5.1776	5.4569	< 0.001
rs2427254,rs13042941	Chr. 20: 60752426 - 62808862	*ADRM1*	4.6456	4.8991	< 0.001
		HRH3	4.6456	4.8991	< 0.001
		LSM14B	4.6456	4.8991	< 0.001
		MIR1257	4.6456	4.8991	< 0.001
		OSBPL2	4.6456	4.8991	< 0.001
		PSMA7	4.6456	4.8991	< 0.001
		SS18L1	4.6456	4.8991	< 0.001
		TAF4	4.6456	4.8991	< 0.001
		CDH4	4.6456	4.8991	0.002

Note: As per significance criteria, for all significantly associating gene regions presented above, the specified pairwise model was chosen as the best overall model and reported a CVC of 10/10.

[1] This is the total region identified by MOCA after combining the gene regions in which the specified model was statistically significant

[2] Gene regions in which the specified qMDR model was statistically significant

[3,4,5] qMDR model statistics calculated for each gene region separately

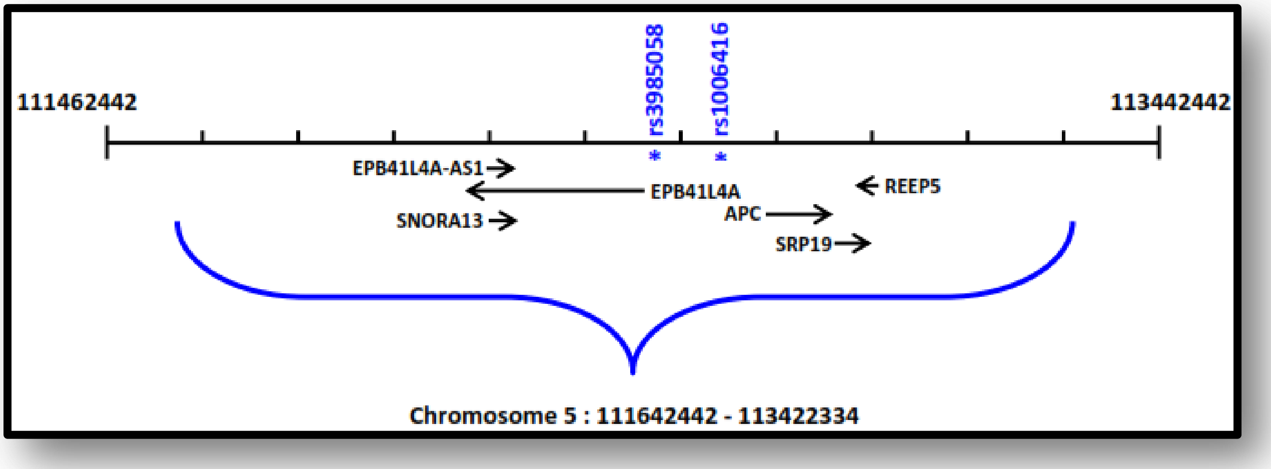
An immediately noticeable trend in Table 4-2 is the identical, or nearly identical, qMDR test statistics for gene regions located within the same locus of interest. This is a result of the mapping technique applied using MOCA that allows SNPs to map to multiple gene regions. This may result in a number of gene regions with highly overlapping and in some cases identical SNP sets. Because qMDR analyses were performed independently in each SNP set (assigned to a gene region) this would result in qMDR reporting identical or very similar model statistics. The statistics shown in the table above indicates that the SNP sets between genes in the same locus are in many instances identical. A complete list of the SNPs assigned to each significantly associated gene region can be found in Appendix Tables 5a-5d.

Evaluation of Pairwise Interaction Models defining Regions of Interest Identified by MOCA

Chromosome 5

The pairwise model containing rs3985058 and rs10064163 (rs3985058 – rs10064163) identified a locus on chromosome 5 spanning a total of 1779.9kb (chr5:111642442 – 113422334) defined by six significantly or marginally associated gene regions (Figure 4-4). These gene regions included the *receptor accessory protein 5 (REEP5), adenomatous polyposis coli (APC), signal recognition particle 19kDa (SRP19), Erythrocyte membrane protein band 4.1 like 4A (EPB41L4A), EPB41L4A antisense RNA 1 (EPB41L4A-AS1)*, and *small nucleolar RNA, H/ACA box 13on chromosome 5 (SNORA13)*. The rs3985058 – rs10064163 model was significantly associated with PAI-1 in five of the six assigned genes (p ≤ 0.001), and marginally associated with PAI-1 in the sixth (*EPB41L4A,* p = 0.002)(Table 4-2).

Figure 4-4. Region of Interest on Chromosome 5 containing possible novel PAI-1 candidate genes defined by the rs3985058 – rs10064163 interaction model

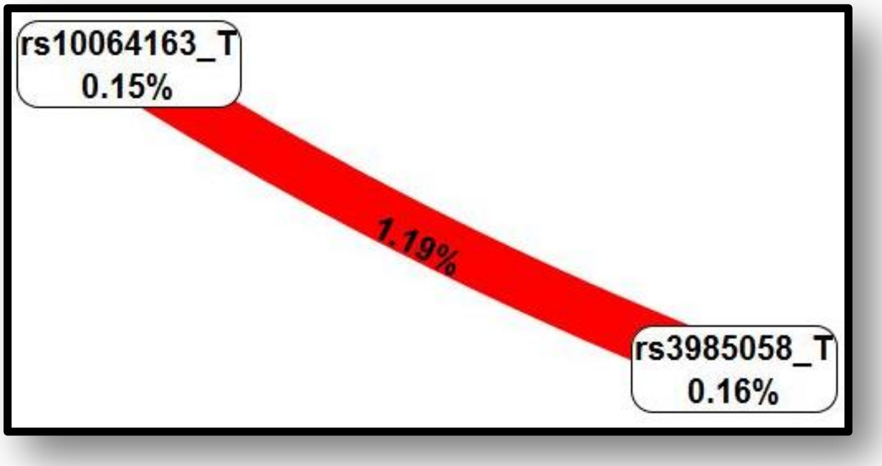


Arrows indicate direction of gene transcription, blue lines indicate boundaries of region of interest. SNPs in the pairwise model are shown in blue bold print.

Both members of the pairwise model identifying the region of interest on chromosome 5 (rs3985058 – rs10064163) were located outside of formal gene boundaries but were in an intergenic region between *EPB4IL4A* and *APC*. Further examination of the model revealed that the interaction between the two SNPs was highly synergistic with the interaction effect explaining 1.19% of the variation in PAI-1, while rs3985058 and rs10064163 explained only 0.16% and 0.15% of the variance, respectively (Figure 4-5).

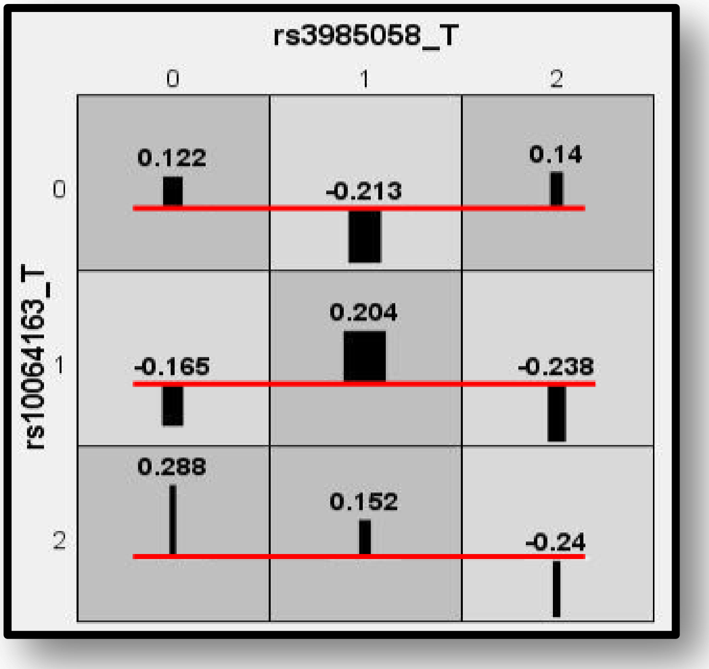
Double heterozygotes displayed the highest mean PAI-1 values (mean = 0.204 ng/mL) (Figure 4-6).Linkage disequilibrium was measured between the two SNPs and they were determined to be unlinked (r^2=0); this information together with the low amount of variance explained by each variant independently provide strong evidence that rs3985058 – rs10064163 is a true epistatic (non-additive) interaction.

Figure 4-5. Visualization of Model Entropy for rs3985058 – rs10064163



Percentages inside boxes indicate the percent of variation in PAI-1 explained by the single SNP; percentage inside connecting line indicates the percentage of variance explained when both markers are considered together.

Figure 4-6. QMDR Graphical Presentation of rs3985058 – rs10064163.



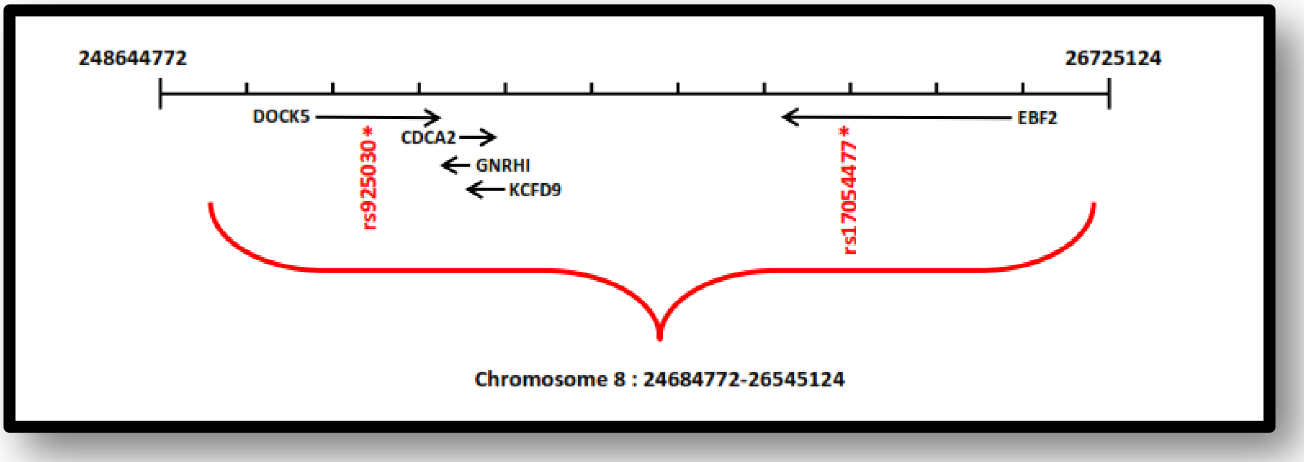
Red line indicates the global average of PAI-1 measurements, regardless of genotype

Numbers and bar height show the difference between the global average of PAI-1 and the genotype-specific average of PAI-2 for a give genotype group. Bar width indicates the proportion of the sample that falls within the specified genotype group. Underscores after SNP name indicate the minor allele at the specified variant.

Chromosome 8

A 1860.4kb region on chromosome 8 was identified by the rs925030 – rs17054477 interaction effect. The locus of interest on chromosome 8 contained 5 gene regions; four of which reported significant associations between the rs925030 – rs17054477 model and PAI-1 levels, the exception was the *dedicator of cytokinesis 5 gene (DOCK5)* (p=0.003). The other gene regions included in this region were the *cell division cycle associated 2* (*CDCA2)*, *early B-cell factor 2 (EBF2)*, *gonadotropin-releasing hormone 1 (GNHR1), and potassium channel tetramerization domain containing 9 (KCTD9).*

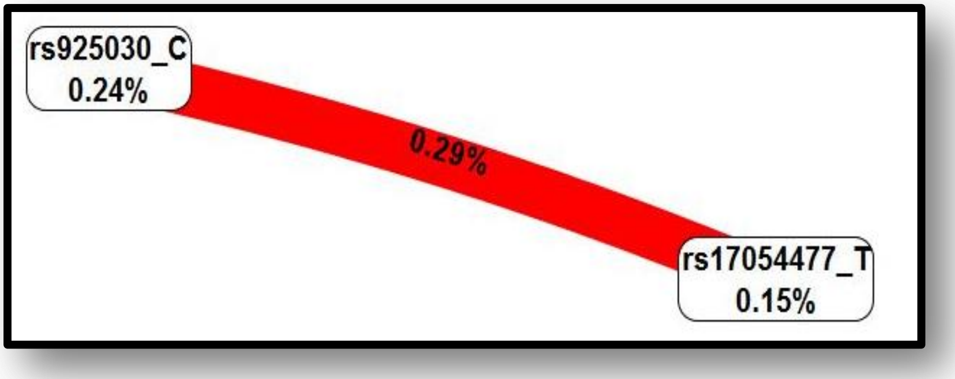
Figure 4-7. Region of interest on Chromosome 8 identified by MOCA



Arrows indicate direction of gene transcription, red lines indicate boundaries of region of interest. SNPs in the pairwise model are shown in blue bold print.

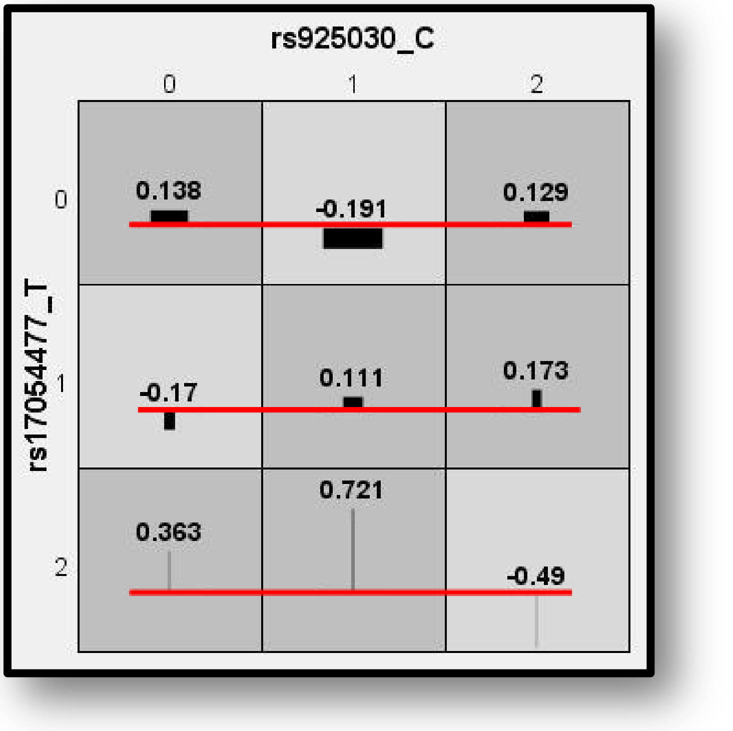
In contrast to the pairwise model on chromosome 5, both rs17054477 and rs925030 are located inside genes within the locus of interest on chromosome 8. Rs17054477 is a missense SNP located in exon 15 of *EBF2* that causes a glycine to serine shift. Additionally, rs17054477 is located proximal to a predicted exon splice enhancer site; though the variant is predicted to be benign by PolyPhen. Rs925030 is located in an intronic region inside *DOCK5*, and to our knowledge has no predicted functional effects. The interaction between rs17054477 was synergistic in nature (percentage of variance in PAI-1 explained by rs17054477 – rs925030 = 0.29%) as depicted in Figure 4-8. LD was measured between the two SNPs to determine if they each represented an independent signal, and testing showed that rs925030 and rs17054477 were in complete linkage equilibrium ($r^2 = 0$). Further evaluation of the model revealed that the largest genotypic group (individuals heterozygous at rs925030 and homozygous for the major allele (C allele) at rs17054477) had decreased PAI-1 as compared to the global mean. The most exaggerated decrease in mean PAI-1 compared to the global mean, and also to all other genotype groups was seen in subjects who were homozygous for both minor alleles; however, this should be interpreted with caution as there were only three individuals in that genotypic category.

Figure 4-8. Visualization of Model Entropy for rs925030 – rs17054477



Percentages inside boxes indicate the percent of variation in PAI-1 explained by the single SNP; percentage inside connecting line indicates the percentage of variance explained when both markers are considered together.

Figure 4-9. QMDR Graphical Presentation of rs962050 – rs1705447



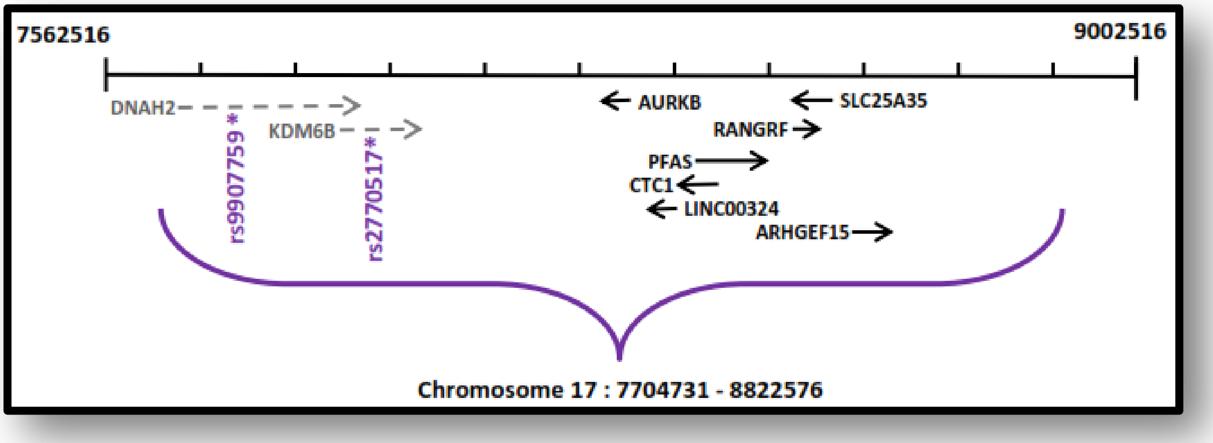
Red line indicates the global average of PAI-1 measurements, regardless of genotype

Numbers and bar height show the difference between the global average of PAI-1 and the genotype-specific average of PAI-2 for a give genotype group. Bar width indicates the proportion of the sample that falls within the specified genotype group. Underscores after SNP name indicate the minor allele at the specified variant.

Chromosome 17

The significant pairwise interaction between rs9907759 and rs2770517 pinpointed a 1117.8 kb region on chromosome 17 that included seven significantly associating genes. These genes were *Rho guanine nucleotide exchange factor (ARHGEF15), aurora kinase B (AURKB), CTS telomere maintenance complex component 1 (CTC1), long intergenic non-protein coding RNA 324 (LINC00324), phosphoribosylformylglycinamidine synthase (PFAS), RAN guanine nucleotide release factor (RANGRF),* and *solute carrier family 25, member 35 (SLC25A35)*. In contrast to results in previous regions, while rs9907759 and rs2770517 were both located within genic regions inside the locus of interest on chromosome 17, neither was the single best model for the genes in which they reside. Rs9907750 is located in an intronic region of the *dynein, axonemal, heavy chain 2* (*DNAH2*) gene, while rs2770517 is located in an intronic region of the *lysine (K)-specific demethylase 6B (KDM6B)* gene. Interestingly, qMDR analyses in both *DNAH2* and *KDM6B* reported multiple “best models”, including the rs9907759 – rs2770517 interaction effect. This indicates that both of these regions may be harboring several pairwise effects of the same intensity and qMDR was unable to classify any one model as “best”. In instances such as this, qMDR will randomly select one of the “best” models and use this to calculate model statistics (T-statistics, CVC, pvalue). In the *KDM6B* gene region, the rs9907759 – rs2770517 interaction reported a CVC of 7/10, and permutation p-value = 0.085; and in *DNAH2* the model showed a CVC of 6/10 and a p-value of 0.056.

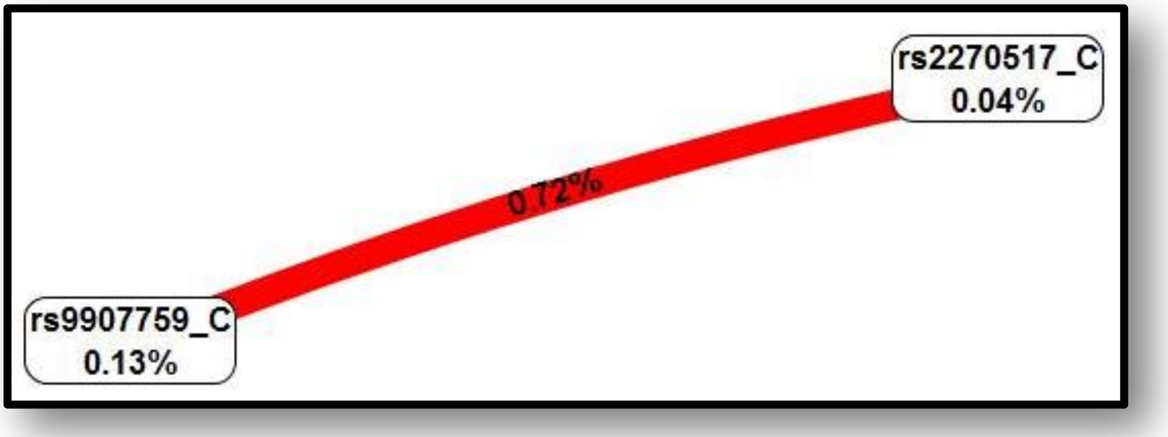
Figure 4-10. Region of interest on Chromosome 17 Identified by MOCA.



Arrows indicate direction of gene transcription, purple lines indicate boundaries of region of interest. Dotted line arrows denote genes in which the rs99007759 – rs2770517 was not significantly associated with PAI-1. SNPs in the pairwise model are shown in purple bold print.

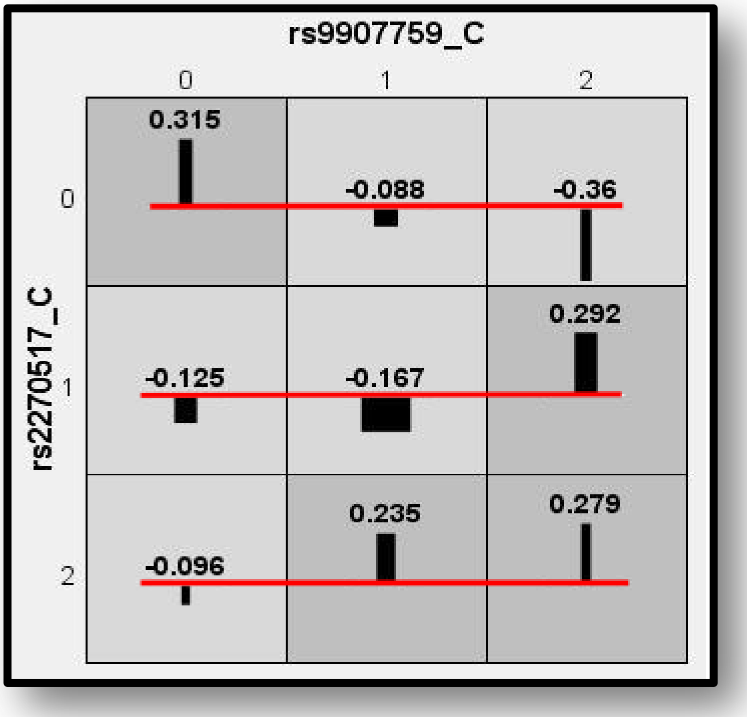
LD analysis to measure the correlation between rs9907759 and rs2270517 was performed and confirmed that the two SNPs were not in linkage disequilibrium with each other (r^2 = 0) and provided further evidence that the observed multi-locus effect between them was truly an interaction effect. The nature of the rs9907759 – rs2270517 effect, as those seen in chromosomes 5 and 8, was found to be synergistic, explaining 0.72% of the variation in PAI-1 while individually rs9907759 and rs2270517 explained only 0.13% and 0.04%, respectively (Figure 4-10). The highest PAI-1 measurements, compared to the global average of PAI-1, were seen in double homozygotes (major allele) as shown in Figure 4-11.

Figure 4-11. Visualization of Model Entropy for rs2270517 – rs9907759



Percentages inside boxes indicate the percent of variation in PAI-1 explained by the single SNP; percentage inside connecting line indicates the percentage of variance explained when both markers are considered together.

Figure 4-12. QMDR Graphical Presentation of rs2270517 – rs9907759



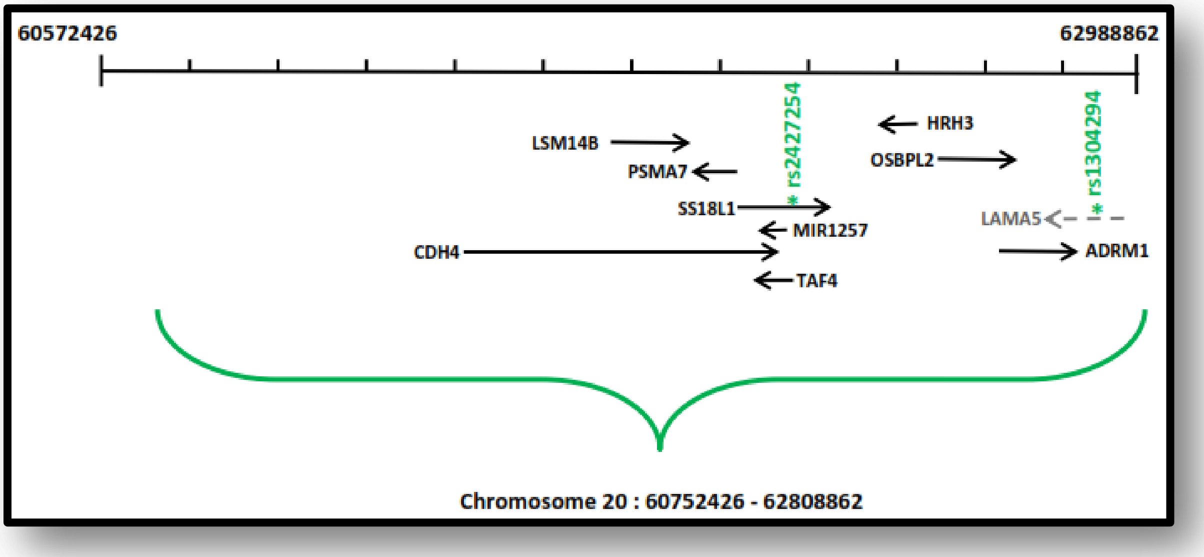
Red line indicates the global average of PAI-1 measurements, regardless of genotype

Numbers and bar height show the difference between the global average of PAI-1 and the genotype-specific average of PAI-2 for a give genotype group. Bar width indicates the proportion of the sample that falls within the specified genotype group. Underscores after SNP name indicate the minor allele at the specified variant.

Chromosome 20

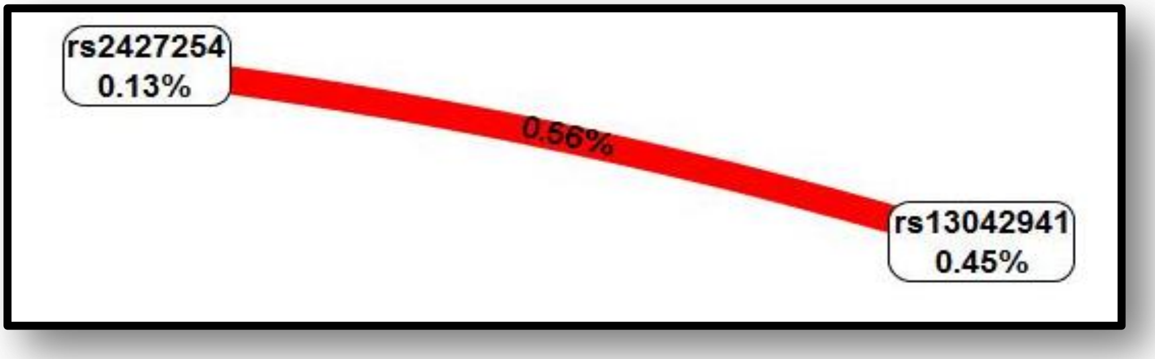
The last locus of interest identified by MOCA was a 2056.4kb region on chromosome 20 that was localized via the significant pairwise interaction between rs2427254 and rs13042941 (rs2427254 – rs13042941). Containing nine novel candidate genes, the locus identified on chromosome 20 is the largest of the regions identified through the MOCA pipeline. The genes found in this region included the *adhesion regulation molecule 1 (ADRM1)*, *histamine receptor H3 (HRH3)*, the *LSM14B, SCD6 homolog B (LSM14B)*, *microRNA 1257 (MIR1257*), *osysterol binding protein-like 2 (OSBPL2)*, *proteasome subunit, alpha type 7 (PSMA7), synovial sarcoma translocation, SS18L1, TAF 4 RNA Polymerase II, TATA Box binding protein associated factor (TAF4)*, and *cadherin 4, type 1, R-cadherin (CDH4)*. The rs2427254 – rs13042941 model was significantly associated with PAI-1 in all the aforementioned genes with the exception of *CDH4* (p = 0.0002) (Table 4-2). Rs2427254 is located in an intronic region of *SS18L1,* while, as seen in the previous model, rs13042941 is not located in any of the genes that identified rs2427254 – rs13042941 as the single "best" model. Rs13042941 is a missense SNP that causes a threonine to alanine shift, located in the *laminin alpha 5* (*LAMA5*) gene. This gene identified multiple models as the "best" model, and rs2427254 – rs13042941 was among these; however the model was not significantly associated with PAI-1 levels in this gene region (CVC = 8/10, p = 0.012) when evaluated using our conservative significance threshold (CVC = 10/10, p <0.001).

Figure 4-13. Region of interest on Chromosome 20 identified by MOCA.



Arrows indicate direction of gene transcription, green lines indicate boundaries of region of interest. Dotted line arrows denote genes in which the rs2427254 – rs13042941 was not significantly associated with PAI-1. SNPs in the pairwise model are shown in purple bold print.

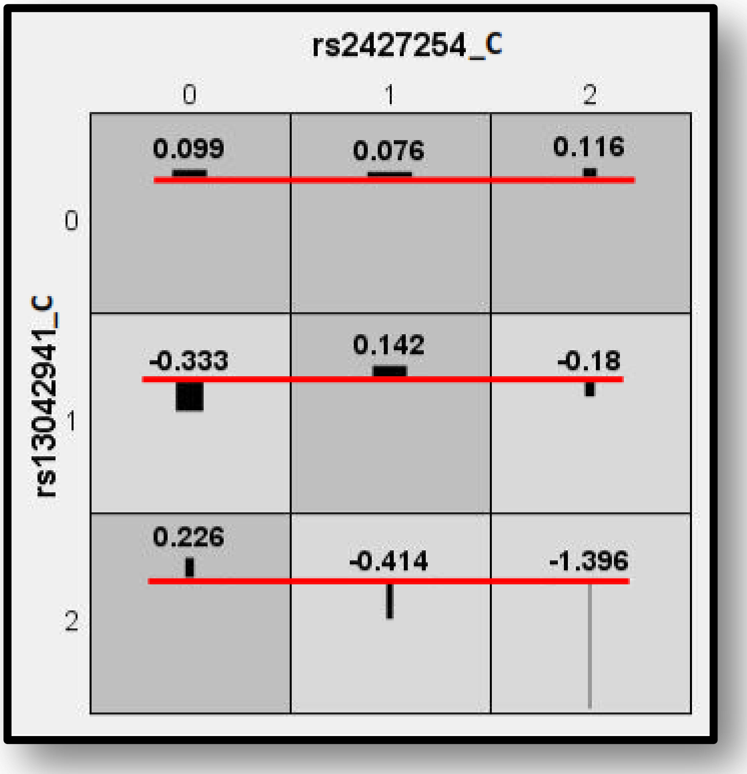
Figure 4-14. Visualization of Model Entropy for rs2427254 – rs13042941



Percentages inside boxes indicate the percent of variation in PAI-1 explained by the single SNP; percentage inside connecting line indicates the percentage of variance explained when both markers are considered together.

QMDR analysis revealed that the rs2427254 – rs13042941 model effect was highly synergistic, with the interaction explaining 0.56% of the variance in PAI-1 levels as seen if Figure 4-14. RS13042941 displayed the highest main effects (0.45% of the variance in PAI-1 was explained by this SNP in the rs2427254 – 13042941 model). LD analyses were performed to determine if the two SNPs in the model were acting independently of each other, and tests confirmed that the two SNPs were not in LD (r^2=0). The largest difference between the genotype group specific means and the global mean was seen in individuals who were heterozygous at rs13042941 and homozygous major at rs2427254; this group displayed lower mean PAI-1 measurements than the global average as depicted in Figure 4-15 below.

Figure 4-15. QMDR Graphical Presentation of rs2427254 – 13042941..



Red line indicates the global average of PAI-1 measurements, regardless of genotype

Numbers and bar height show the difference between the global average of PAI-1 and the genotype-specific average of PAI-2 for a give genotype group. Bar width indicates the proportion of the sample that falls within the specified genotype group. Underscores after SNP name indicate the minor allele at the specified variant.

Discussion

We presented a novel approach that incorporated multi-locus association analyses through the use of qMDR to prioritize novel candidate genes for further analyses of PAI-1 variation as an important step in addition to single SNP analyses in our data. Using this approach we were able to identify four regions of interest on chromosomes 5, 8,17, and 20 and select 28 candidate genes based on the association results from four pairwise SNP-SNP models. This approach represents a much more inclusive method for prioritizing genes as compared to methods that rely on single variant association analyses such as those performed in Chapter III. One caveat of this approach is that as with linkage analyses to identify QTL, MOCA highlights regions of interest not necessarily single genes. Because multiple gene regions can report the same multi-locus signal it is impossible to discern, without performing functional studies, which gene(s) within a region of interest is truly responsible for the association that is being detected. However, unlike genome-wide linkage peaks which can span 1-2cm and contain 100's of genes as well as immense "gene deserts", our method produces smaller gene-centered regions, with the majority of the SNPs being able to be assigned to one or a few genes, making interpretation of results slightly more straightforward.

GWAS, candidate gene, and QTL mapping have all suffered from lack of replication of effects in the context of various cardiovascular disease phenotypes. After identifying loci of interest using MOCA, we ascertained the location of previously identified QTL loci of relevant CVD/PAI-1 associated traits to determine if there was any significant overlap in identified regions. We discovered that all four of the regions identified by MOCA were located under QTL peaks for various CVD and PAI-1 related traits such as triglycerides, Type 2 diabetes, obesity, glucose, and early onset myocardial infarction[146-158]. A summary of these results are depicted in Table 4-3.

Table 4-3. MOCA identified regions of interest that overlap with previously identified CVD / PAI-1 associated Quantitative Trait Loci (QTL).

		Previously Identified Quantitative Trait Loci for Cardiovascular Disease / or PAI-1 related Traits				
	MOCA Region of Interest	QTL Region of Interest	QTL Trait	LOD Score[1]	P[2]	Reference
Chromosome 5	111642442 - 113422334	101218797 - 127218797	*Triglycerides*	1.59	-	Imperatore *et al.*, 2000[142]
		106101774 – 132101774	*Body Mass Index*	1.5	$3.9E^{-3}$	Feitosa *et al.,* 2002[143]
		106101774 - 132101774	*early onset myocardial infarction*	-	-	Wang *et al.,* 2002[144]
Chromosome 8	24684772-26545124	21451041 - 47451041	*Body Weight*	1.3	-	Wilson, SG *et al.,* 2006[145]
		22433395 - 38326352	*HDL Cholesterol*	1.2	-	Arya, R *et al.,* 2002[146]
		25441235 - 106103348	*adiponectin level*	1.6	-	Tejero, NE *et al,.* 2007[147]
Chromosome 17	7704731 - 8822516	432538 - 13171261	*Body Mass Index*	2.25	$7.0E^{-4}$	Meyre, D *et al.*, 2004[148]
		1 - 13623317	*subcutaneous body fat*	1.86	-	Norris, JM *et al.*, 2005[149]
		1260796 - 27260796	*glucose level*	3.74	$1.7E^{-5}$	Wiltshire, S *et al.*, 2001[150]
Chromosome 20	60752426 - 62808862	44053450 - 700535450	*triglycerides*	2.34		Li, ND *et al.,* 2005[151]
		39168716 - 65168716	*Body Weight*	3.16	$6.9E^{-5}$	Lee, HJ *et al.,* 1999[152]
		36561838 - 35995245	*Type 2 Diabetes*	2.92	$6.5E^{-4}$	Luo, TH *et al., 2003*[153]

<u>*Note:*</u> Above Information on QTL position and Study Statistics was obtained via the Human Genome Browser of the Rat Genome Database[158]

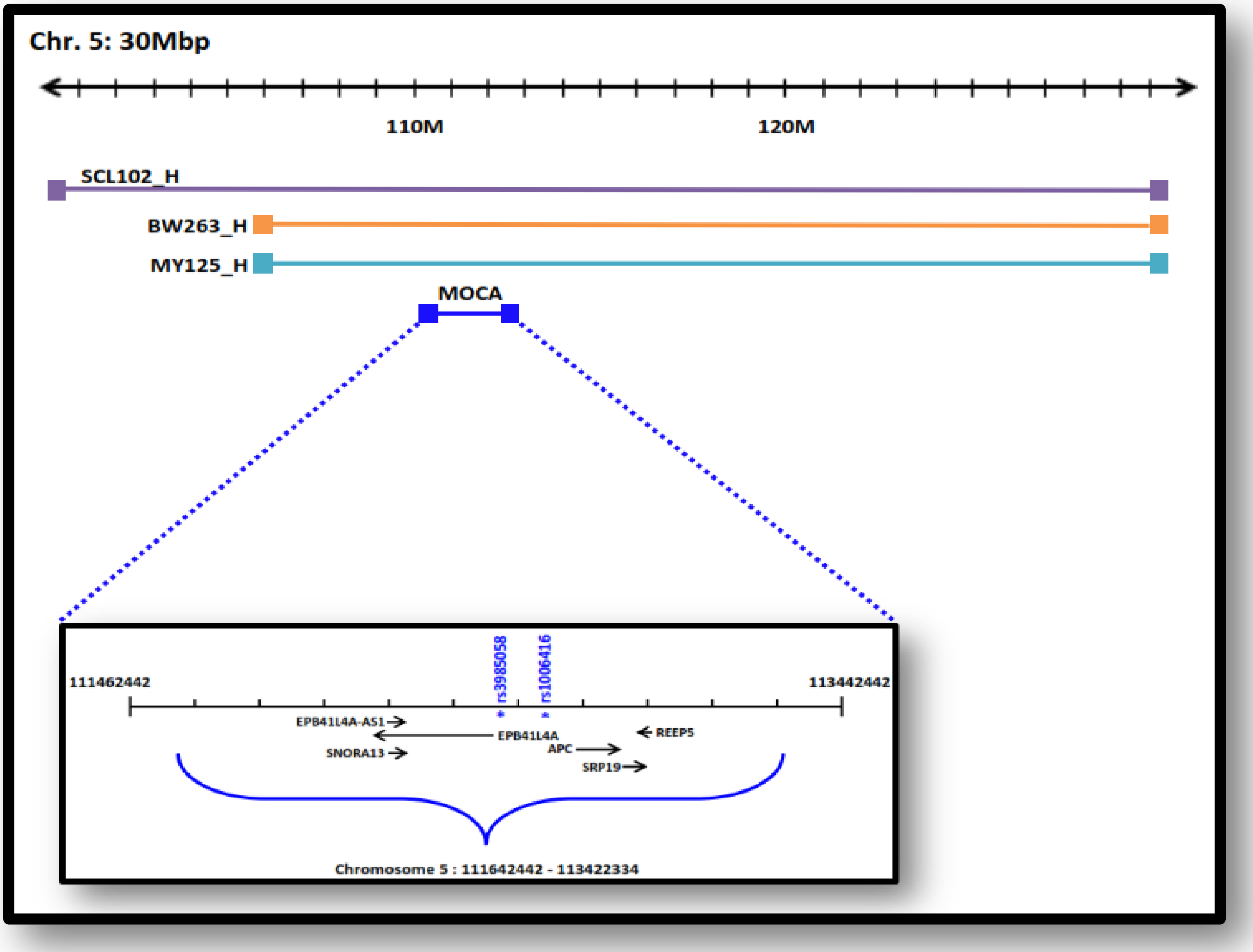
[1] LOD score reported for QTL from specified study

[2] P-value associated with LOD Score for specified study

A review of the gene regions contained within our MOCA selected regions of interest revealed several genes with intuitive biological connections with both PAI-1 and each other. As an example of the utility and efficacy of our method in prioritizing genes with plausible mechanism in connection with the variation in PAI-1 levels provide a preliminary examination of the interconnections between several genes identified by MOCA inside the loci of interest located on Chromosome 5. This region was chosen to highlight the efficacy of MOCA because the MOCA identified region fell within three

QTL regions for three traits that have been shown in previous studies in humans and animal models to associate with PAI-1 levels; body mass index, early onset myocardial infarction, and triglycerides (Figure 4-16).

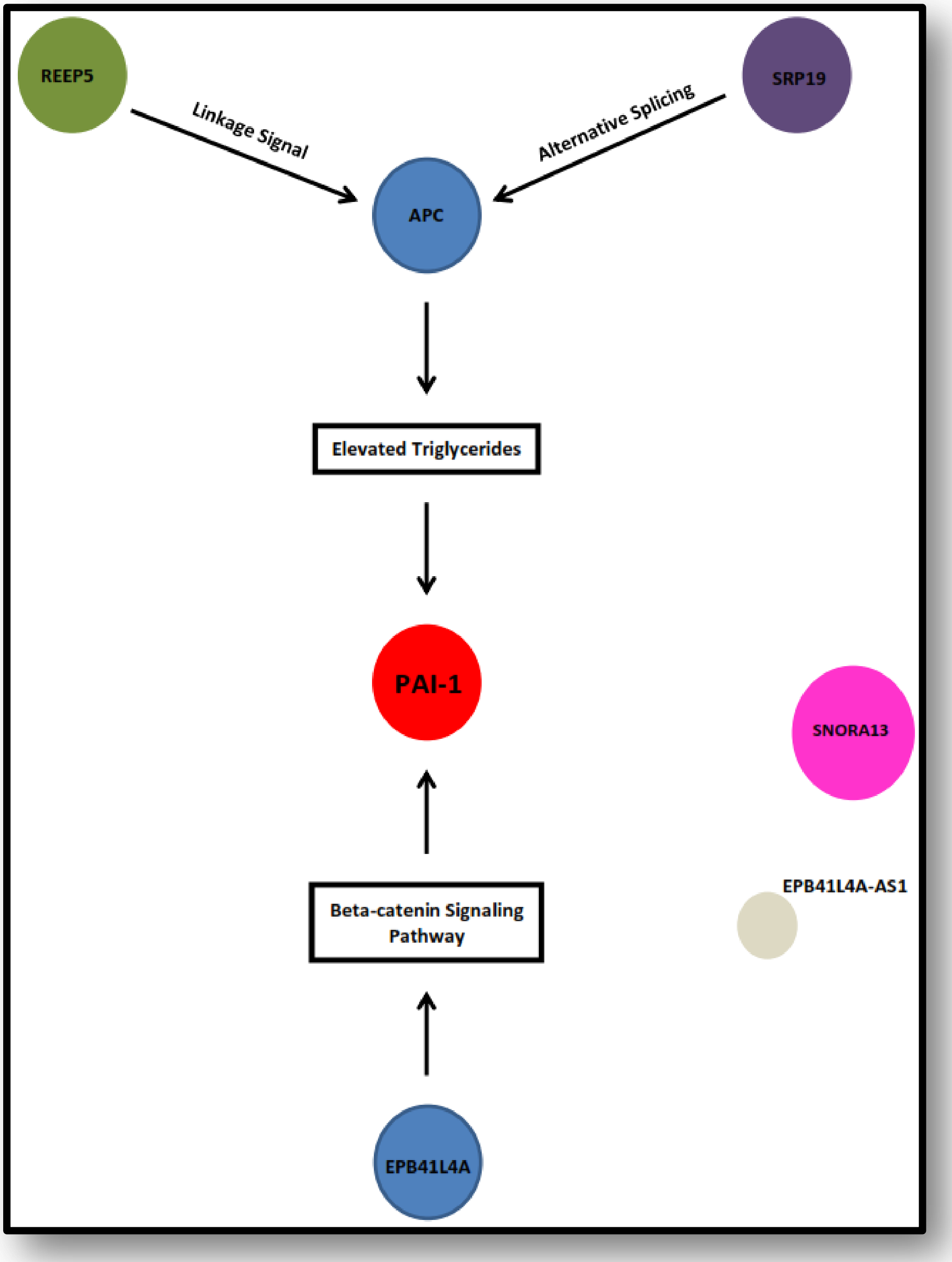
Figure 4-16. Chromosome 5 MOCA Identified Region of Interest Compared to Previously Identified QTL Regions.



Note: QTL regions are defined using Rat Genome Database Nomenclature.
SCL102_H: Human Triglycerides QTL
BW263_H: Human Body Mass Index QTL
MY125_H: Human Early-onset Myocardial Infarction QTL

To showcase the efficiency of the MOCA approach we will evaluate the possible interconnections between the genes in the chromosome 5 region as this loci fall under the most clinically relevant CVD/PAI-1 QTL loci. *REEP5* was discovered in 1991 to be tightly linked to the *APC* gene, and is used in predictive *APC* mutation carrier screening for familial adenomatous polyposis (FAP)[159,160]. Multiple forms of *APC* RNA transcripts have been discovered in various tissues in humans and model organisms; one of these transcripts has been shown to undergo an alternative splicing event at exon 14 that leads to the incorporation of the *SRP19* gene[161]. After establishing the connection of *REEP5* and *SRP19* to *APC*, we then identified a connection between APC and PAI-1 via triglyceride levels. Studies in model organisms indicate a strong connection between variation in *APC* and high serum triglyceride levels; elevated triglycerides are a known biological modulator of PAI-1 activity[39,162]. Another plausible connection between APC and PAI-1 is through the cyclic-AMP/cAMP response element binding protein (cAMP/CREB) signaling pathway which has been shown to target PAI-1; the cAMP/CREB pathway has been shown to induce *PAI-1* expression in the liver under various conditions and *CREB* itself has the ability to bind to the *PAI-1* promoter and induce expression of PAI-1 under oxidative conditions[163-165]. Another plausible connection to PAI-1 was found with EPB4IL4A through its participation in the beta-catenin signaling pathway, of which *PAI-1* is a direct transcriptional target; activation of beta-catenin induces PAI-l expression and promoter activity[166]. *EPB41L4A*, previously referred to as *NBL4*, is an important component of the beta-catenin pathway; additionally, there is also evidence that *APC* is a major regulator of this pathway[167,168]. We were unable to discern intuitive connections between SNORA13 or EPB41L4a-AS1 and PAI-1, but it is possible that the biological relationship between these two genes and PAI-1 is more subtle, and may be revealed under more direct scrutiny. A visualization of these plausible biological and mechanistic connections between and among the genes in the chromosome 5 loci of interest and PAI-1 is presented in Figure 4-16.

Figure 4-17. Visualization of Possible Biological connections between genes in the Chromosome 5 Region of Interest and PAI-1



Genes are indicated by adjacent or encompassing circles, and arrows indicate possible directional relationships

CHAPTER V

CONCLUSIONS AND FUTURE DIRECTIONS

A. Summary

Cardiovascular disease (CVD) is a leading cause of mortality both in the United States and globally. CVD is a complex disease that describes several independent but etiologically overlapping disorders that are likely to be heavily influenced by both genetic and environmental factors. Major thrombotic events due in part to decreased or imbalanced fibrinolysis are a unifying characteristic among several major CVD's and understanding the genetic factors that impact the fibrinolytic system has the potential to reveal information that may be applied to several CVDs. Utilizing PAI-1 levels as an endophenotype in the investigation of the genetic impact on CVD susceptibility has the advantage of revealing information that is more likely to be universal in its application to CVD and less affected by the specific disease etiology of any one cardiovascular phenotype. Although PAI-1 has been assessed as a biomarker of both fibrinolytic activity as well as CVD susceptibility and severity, the majority of the heritability of this potent biomarker remains unexplained. This issue is particularly evident in non-Europeans, where the few studies that have been done to date have been candidate gene studies that interrogated only a small number of variants, and may not have used the most appropriate methods to analyze the genetics of PAI-1. Studies presented in the previous chapters were aimed at elucidating the genetic architecture underlying PAI-1 variation in the current study and directing future research in the continued evaluation of PAI-1 as an endophenotype for CVD.

There are several important points that were revealed by the studies presented in this dissertation. First, while many studies have assessed the impact of single variants on variation in PAI-1 levels in several populations, many of these are inconsistent. These inconsistencies may be due in part to incorrect statistical inference due to the use of parametric methods that were inappropriate in the context of their respective study populations. We have demonstrated that when the characteristic non-normality of the PAI-1 distribution is not taken into account, there is a significant impact on the statistical results and corresponding inference obtained from applying standard parametric methods. The use of inappropriate parametric methods in the evaluation of genetic impact on PAI-1 levels may explain, in part, some of the inconsistencies of previous studies regarding single variant associations with PAI-1. Using median regression, a non-parametric method well suited to the PAI-1 distribution observed in the HeART cohort, we not only revealed novel associations with PAI-1 levels we also provided a direct assessment of the impact of incorrect inference that would have resulted from our use of standard linear regression. This highlights the possibility that significant associations present in other studies may have gone unnoticed.

Secondly, although elevated PAI-1 levels have been shown to associate with increased CVD susceptibility and severity, few studies have conducted a comprehensive evaluation of genetic variation affecting this clinically relevant portion of the PAI-1 distribution. We provide, to our knowledge, the only study that explicitly tests the hypothesis of non-uniform SNP effects within the PAI-1 phenotypic distribution. Our efforts revealed significant associations with the upper quartile of PAI-1 values, but also highlighted the lack of overlap between markers associated with median PAI-1 and those associated with elevated PAI-1 levels. This illustrated for the first time that variants that impact PAI-1 levels at the upper extreme of the distribution may have only nominal impact on the distribution as a whole.

Thirdly, our association analyses in the upper quartile of PAI-1 revealed an association with PER3, an important regulator of the circadian clock pathway. This result was particularly poignant for two reasons; the first is that studies in animal models revealed a direct functional connection between the CLOCK-ARNTL heterodimer and activation of the PAI-1 promoter[81], and the second was that a recent meta-analysis in several Caucasian populations also identified ARNTL as significantly associating with PAI-1. The interplay between PER3 and ARNTL is essential to the proper function of the Circadian clock pathway and the identification of two genes from the same pathway in different populations provides supporting evidence that there may be a generalizable Circadian pathway effect on variation in PAI-l, although specific gene-level effects within the pathway may be population specific.

Lastly, we presented a novel approach to prioritize the selection of candidate genes for further evaluation of genetic effects on PAI-1 as a necessary and logical next step to our single variant level analyses. We showed the utility of using multi-locus association signals to localize regions containing genes that not only contain evidence of variation associated with PAI-1 but are also likely to have intuitive biological connections with PAI-1 as well. We show the overlap between the regions identified by our approach, MOCA, and those identified by QTL mapping of cardiovascular disease and / or PAI-1 related traits/biomarkers. We also demonstrated the utility of our method not only in the selection of candidate genes for future studies but also in the detection of significantly associating pairwise interactions

B. Future Directions

PAI-1 is a potent biomarker, known to be affected by both environmental and genetic factors, that is involved in a complex web of numerous biological processes in addition to its essential role in the fibrinolytic system. Our genetic association studies of PAI-1, while revealing novel genetic effects on variation in PAI-1 with evidence of plausible biological connections to PAI-1 expression, we were unable to test the integrity of our finding through replication in a similar population. We chose to study PAI-1 variation in a West African population to address a striking deficiency in the current knowledge base on variation affecting this potent biomarker of CVD. We selected as our study sample all urban participants of the HeART cohort with available 4G/5G genotype data. As previously mentioned, the HeART cohort is one of the largest population-based African cohorts with available PAI-1 measurements; unfortunately it is also one of the only sources of information regarding PAI-1 variation in Africa. Due to this fact we were unable to identify an appropriate replication cohort in order to validate the significant association effects that we uncovered in our studies.

Therefore, while these studies have identified genes of interest, replication in a similar cohort is needed to confirm the validity of these findings. Additionally, our assessed markers overlap with variants assessed in GWAS of PAI-1 variation in non-African populations but the majority of associations that we report are unique; this fact is particularly troubling in reference to our significant associations detected using median regression analyses, as median regression analyses are analogous to linear regression when model assumptions are upheld. This inconsistency may be due to population-specific effects, false-positive results in our study, or could be a result of incorrect inference in previously studied cohorts that inhibited detection of these associations. Additional accompanying studies in non-African populations employing the non-parametric methods outlined in this dissertation, where appropriate, are required to address these concerns, as well as assess the generalizability across multiple populations of the significant associations that we have identified.

The identification of supporting evidence for a possible pathway-level effect regarding the Circadian clock pathway, through the discovery of a significant association with *PER3* revealed through upper quartile regression analyses, highlights the possible importance of pathway level effects on variation in PAI-1. There is evidence that effects at the pathway level may be more likely to generalize between populations as they may be more robust to differences in underlying genetic structure between populations, such as allele frequency differences and LD patterns. A thorough investigation of pathway level effects in several populations may lead to the discovery of universal effects on PAI-1 variation. This will require the use of analytical techniques that can incorporate differences in LD structure between populations, in the analysis of pathway based effects.

As a part of our studies we presented, MOCA, a novel use of qMDR to select candidate genes based on multi-locus association signals. While our identified regions of interest fell within previously discovered QTL for CVD/PAI-1 related traits, and we were able to draw intuitive connections between our identified genes and PAI-1, we did not formally assess the power of this approach to prioritize genes versus those employed in GWAS and candidate gene study designs that rely on single variant association signals. Follow-up studies, perhaps involving simulation studies, are needed to formally assess the power of this approach under different underlying genetic models to determine if our results from using the approach are generalizable.

The identification of significantly associated synergistic interaction effects highlights the need for formal evaluation of gene-gene and gene-environment associations in the investigation of PAI-1 variation. Our studies revealed pairwise effects in the absence of strongly associating main effects; this highlights the possibility that higher-order effects may have a significant impact on variation in PAI-1 levels that may have been overlooked in previous studies. Another avenue worthy of further evaluation would be the investigation of haplotype effects in the regions identified via MOCA analysis to identify any significant haplotype effects that may impact PAI-1 levels. Sliding window haplotypes of all single variants within these regions may reveal novel inter- and intragenic haplotype effects. A truly comprehensive understanding of the variation in PAI-1 levels requires careful evaluation of pairwise and higher-order interaction effects, and the implementation of analytical methods that acknowledge the complexity of this CVD endophenotype.

Appendix Table 1. Hardy-Weinberg Equilibrium Estimates and allele frequencies of SNPs significantly associated with Median Plasminogen Activator Inhibitor-1 (PAI-1) levels

Chr.	Gene	SNP	Minor Allele	Major Allele	MAF[1]	HWE P-value[2]
5	*ARSB*	rs1071598	T	C	0.048	0.726
7	*CPA2*	rs61997065	A	G	0.045	0.466
19	*LENG9*	rs10406453	T	C	0.075	0.652

[1] MAF; Minor Allele Frequency
[2] HWE P-value; Hardy-Weinberg Equilibrium P-value

Appendix Table 2. Genotypic Distribution of SNPs significantly associated with Median Plasminogen Activator Inhibitor 1 (PAI-1) levels

Chr.	Gene	SNP[1]	Minor Allele	Major Allele	Genotype Distribution[2]		
					mm	*Mm*	*MM*
5	*ARSB*	rs1071598	T	C	1	98	954
		rs1071598_dom			99		954
7	*CPA2*	rs61997065	A	G	3	89	961
		rs61997065_dom			92		961
19	*LENG9*	rs10406453	T	C	7	143	902

[1] Instances in which sample size was below 5 for any genoptype group, SNPs were recoded dominantly for the effect of the minor allele (homozygous minor and heterozygotes were combined) prior to regression analyses; *_dom* denotes dominant coding genotype distribution
[2] *mm* = homozygous minor, *Mm* = heterozygote, *MM* = homozygous major

Appendix Table 3. Hardy-Weinberg Equilibrium Estimates and allele frequencies of SNPS significantly associated with the Upper Quartile of Plasminogen Activator Inhibitor-1 (PAI-1) Distribution

Chr.	Gene	SNP	Minor Allele	Major Allele	MAF[1]	HWE P-value[2]
1	*COL16A1*	rs72887331	A	C	0.141	0.610
1	*FHAD1*	rs12126178	A	G	0.131	0.177
1	*PER3*	rs10462021	G	A	0.070	0.473
2	*PLECKHB2*	rs6713972	G	T	0.088	0.029
3	--	rs13314993	T	G	0.077	0.272
3	*SLC15A2*	rs116307792	G	A	0.054	1.000
5	*ADAMTS12*	rs61757473	C	G	0.049	0.509
6	*TAGAP*	rs35263580	T	C	0.053	0.113
7	--	rs2023783	A	G	0.070	0.475
9	*DBH*	rs4531	T	G	0.146	0.901
11	*EXT2*	rs4755779	G	A	0.071	0.231
11	*PHLDB1 / TREH*	rs7389	C	A	0.232	0.339
	TREH	rs519982	T	C	0.230	0.163
12	*OR1OP1*	rs76940436	T	A	0.065	0.441
12	*P2RX7*	rs34219304	A	G	0.050	0.177
14	*NID2*	rs2273430	C	A	0.248	0.868
14	*FAM161B*	rs34834232	T	A	0.114	0.359
16	*C1QTNF8*	rs73494080	G	T	0.051	0.105
17	*CEP95*	rs9910506	A	G	0.055	1.000

MAF; Minor Allele Frequency

[2] HWE P-value; Hardy-Weinberg Equilibrium P-value

Appendix Table 4. Genotypic Distribution of SNPs significantly associated with the Upper Quartile of the Plasminogen Activator Inhibitor 1 (PAI-1) Distribution

Chr.	Gene	SNP[1]	Minor Allele	Major Allele	Genotype Distritution[2]		
					mm	*Mm*	*MM*
1	*COL16A1*	rs72887331	A	C	23	251	779
1	*FHAD1*	rs12126178	A	G	23	230	800
1	*PER3*	rs10462021	G	A	3	141	909
		rs10462021_*dom*			144		909
2	*PLECKHB2*	rs6713972	G	T	14	153	867
3	--	rs13314993	T	G	9	144	900
3	*SLC15A2*	rs116307792	G	A	3	108	942
5	*ADAMTS12*	rs61757473					
6	*TAGAP*	rs35263580	T	C	6	100	944
		rs35263580_*dom*			106		944
7	--	rs2023783	A	G	3	142	908
		rs2023783_*dom*			145		908
9	*DBH*	rs4531	T	G	23	261	769
11	*EXT2*	rs4755779	G	A	8	132	913
11	*PHLDB1 / TREH*	rs7389	C	A	50	384	610
	TREH	rs519982	T	C	47	390	616
12	*OR1OP1*	rs76940436	T	A	6	125	922
12	*P2RX7*	rs34219304	A	G	5	95	952
14	*FAM161B*	rs34834232	T	A	17	207	829
14	*NID2*	rs2273430	C	A	65	387	591
16	*C1QTNF8*	rs73494080	G	T	0	107	946
		rs73494080_*dom*			107		946
17	*CEP95*	rs9910506	A	G	3	110	940
		rs9910506_*dom*			113		940
20	*DEFB132*	rs74420259	A	G	3	112	937
		rs74420259_*dom*			115		937

[1] Instances in which sample size was below 5 for any genoptype group, SNPs were recoded dominantly for the effect of the minor allele (homozygous minor and heterozygotes were combined) prior to regression analyses; _*dom* denotes dominant coding genotype distribution

[2] *mm* = homozygous minor, *Mm* = heterozygote, *MM* = homozygous major

Appendix Table 5a. SNPs Mapped to Genes within the Chromosome 5 Region of Interest

Chromosome 5 Genes	*REEP5*	*SRP19*	*APC*	*EPB41L4A-A*	*SNORA13*	*EPB41L4A*
Mapped SNPs	rs3985058	rs3985058	rs890757	rs255888	rs255888	rs255888
	rs10064163	rs10064163	rs3985058	rs34927	rs34927	rs34927
	rs469727	rs469727	rs10064163	rs7703522	rs7703522	rs7703522
	rs9326869	rs9326869	rs469727	rs1560058	rs1560058	rs1560058
	rs4705752	rs4705752	rs9326869	rs7719346	rs7719346	rs7719346
	rs17135515	rs17135515	rs4705752	rs13165201	rs13165201	rs13165201
	rs1318772	rs1318772	rs17135515	rs890757	rs890757	rs890757
				rs3985058	rs3985058	rs3985058
				rs10064163	rs10064163	rs10064163
						rs469727

Appendix Table 5b. SNPs Mapped to Genes within the Chromosome 8 Region of Interest

Chromosome 8 Genes	*CDCA2*	*EBF2*	*GNRH1*	*KCTD9*	*DOCK5*	*CDCA2*
Mapped SNPs	rs35475676	rs925030	rs35475676	rs35475676	rs4871930	rs35475676
	rs17053341	rs115875864	rs17053341	rs17053341	rs196864	rs17053341
	rs925030	rs17054477	rs925030	rs925030	rs35475676	rs925030
	rs115875864	rs2233701	rs115875864	rs115875864	rs17053341	rs115875864
	rs17054477		rs17054477	rs17054477	rs925030	rs17054477
					rs115875864	
					rs17054477	

Appendix Table 5c. SNPs Mapped to Genes within the Chromosome 17 Region of Interest

Chromosome 17 Genes	*ARHGEF15*	*AURKB*	*CTC1*	*LINC00324*	*PFAS*	*SLC25A35*	*RANGRF*
Mapped SNPs	rs9907759	rs12941981	rs374230	rs307627	rs7219467	rs11650083	rs12942745
	rs1105813	rs3744263	rs61737334	rs2543540	rs11868946	rs4792002	rs4233018
	rs1106826	rs34873228	rs114366417	rs16956936	rs9903543	rs4239120	rs9907759
	rs57926692	rs1544724	rs78773193	rs839721	rs8066124	rs28743021	rs1105813
	rs8071598	rs1544725	rs634990	rs7219467	rs4791759	rs8069344	rs1106826
	rs4130668	rs307627	rs1376785	rs11868946	rs11653849	rs145035264	rs57926692
	rs2270517	rs2543540	rs61741130	rs9903543	rs9905906	rs9895916	rs8071598
	rs73233606	rs16956936	rs6495754	rs8066124	rs77431839	rs4792147	rs4130668
	rs8522	rs839721	rs2543540	rs4791759	rs11871543	rs7225107	rs2270517
	rs12453250	rs7219467	rs16956936	rs11653849	rs8070826	rs3027205	rs73233606
	rs7217873	rs11868946	rs839721	rs9905906	rs12942745	rs3027229	rs8522
	rs10852894	rs9903543	rs7219467	rs77431839	rs4233018	rs3027232	rs12453250
	rs9908139	rs8066124	rs11868946	rs11871543	rs9907759	rs1059476	rs7217873
	rs61747003	rs4791759	rs9903543	rs8070826	rs1105813	rs3826543	rs10852894
	rs11078718	rs11653849	rs8066124	rs12942745	rs1106826	rs3027238	rs9908139
	rs35421969	rs9905906	rs4791759	rs4233018	rs57926692	rs78390421	rs61747003
	rs11650083	rs77431839	rs11653849	rs9907759	rs8071598	rs9912921	rs11078718
	rs4792002	rs11871543	rs9905906	rs1105813	rs4130668	rs9891699	rs35421969
	rs4239120	rs8070826	rs77431839	rs1106826	rs2270517	rs11078738	rs11650083
	rs28743021	rs12942745	rs11871543	rs57926692	rs73233606	rs9890841	rs4792002
	rs8069344	rs4233018	rs8070826	rs8071598	rs8522	rs871841	rs4239120
	rs145035264	rs9907759	rs12942745	rs4130668	rs12453250	rs3744647	rs28743021
	rs9895916	rs1105813	rs4233018	rs2270517	rs7217873	rs73250854	rs8069344
	rs4792147	rs1106826	rs9907759	rs73233606	rs10852894	rs12601097	rs145035264
	rs7225107	rs57926692	rs1105813	rs8522	rs9908139	rs12936935	rs9895916
	rs3027205	rs8071598	rs1106826	rs12453250	rs61747003	rs370752	rs4792147
	rs3027229	rs4130668	rs57926692	rs7217873	rs11078718	rs74532943	rs7225107
	rs3027232	rs2270517	rs8071598	rs10852894	rs35421969	rs961	rs3027205
	rs1059476	rs73233606	rs4130668	rs9908139	rs11650083	rs7225835	rs3027229
	rs3826543	rs8522	rs2270517	rs61747003	rs4792002	rs74866427	rs3027232
	rs3027238	rs12453250	rs73233606	rs11078718	rs4239120	rs9893451	rs1059476
	rs78390421	rs7217873	rs8522	rs35421969	rs28743021		rs3826543
	rs9912921	rs10852894	rs12453250	rs11650083	rs8069344		rs3027238
	rs9891699	rs9908139	rs7217873	rs4792002	rs145035264		rs78390421
	rs11078738	rs61747003	rs10852894	rs4239120	rs9895916		rs9912921
	rs9890841	rs11078718	rs9908139	rs28743021	rs4792147		rs9891699
	rs871841	rs35421969	rs61747003	rs8069344	rs7225107		rs11078738
	rs3744647	rs11650083	rs11078718	rs145035264	rs3027205		rs9890841
	rs73250854	rs4792002	rs35421969	rs9895916	rs3027229		rs871841
	rs12601097	rs4239120	rs11650083	rs4792147	rs3027232		rs3744647

	rs12936935	rs28743021	rs4792002	rs7225107	rs1059476		rs73250854
	rs370752	rs8069344	rs4239120	rs3027205	rs3826543		rs12601097
	rs74532943	rs145035264	rs28743021	rs3027229	rs3027238		rs12936935
	rs961	rs9895916	rs8069344	rs3027232	rs78390421		rs370752
	rs7225835	rs4792147	rs145035264	rs1059476	rs9912921		rs74532943
	rs74866427	rs7225107	rs9895916	rs3826543	rs9891699		rs961
	rs9893451	rs3027205	rs4792147	rs3027238	rs11078738		rs7225835
	rs2242373	rs3027229	rs7225107	rs78390421	rs9890841		rs74866427
	rs17854013	rs3027232	rs3027205	rs9912921	rs871841		rs9893451
	rs28446092	rs1059476	rs3027229	rs9891699	rs3744647		
		rs3826543	rs3027232	rs11078738	rs73250854		
		rs3027238	rs1059476	rs9890841	rs12601097		
		rs78390421	rs3826543	rs871841	rs12936935		
		rs9912921	rs3027238	rs3744647	rs370752		
		rs9891699	rs78390421	rs73250854	rs74532943		
		rs11078738	rs9912921	rs12601097	rs961		
		rs9890841	rs9891699	rs12936935	rs7225835		
		rs871841	rs11078738	rs370752	rs74866427		
		rs3744647	rs9890841	rs74532943	rs9893451		
		rs73250854	rs871841	rs961			
		rs12601097	rs3744647	rs7225835			
		rs12936935	rs73250854				
		rs370752	rs12601097				
		rs74532943	rs12936935				
		rs961	rs370752				
		rs7225835	rs74532943				
			rs961				
			rs7225835				
			rs74866427				
			rs9893451				

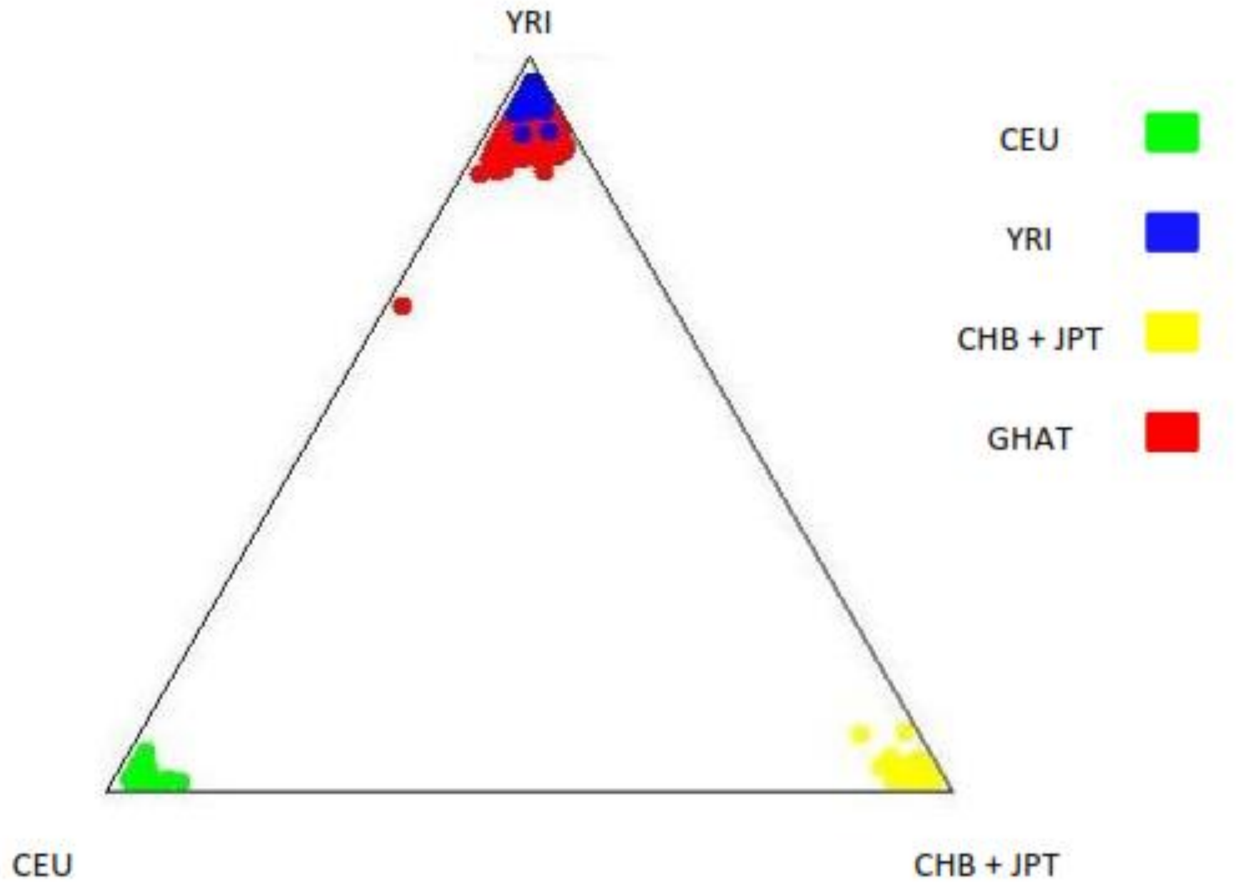
Appendix Table 5d. SNPs Mapped to Genes within the Chromosome 20 Region of Interest

Chromosome 20 Genes	***ADRM1***	***HRH3***	***LSM14B***	***MIRI1257***
Mapped SNPs	rs78979746	rs78979746	rs944260	rs2427158
	rs6142884	rs6142884	rs78979746	rs2024714
	rs2427254	rs2427254	rs6142884	rs944260
	rs36106901	rs36106901	rs2427254	rs78979746
	rs6062133	rs6062133	rs36106901	rs6142884
	rs6142998	rs6142998	rs6062133	rs2427254
	rs77172131	rs77172131	rs6142998	rs36106901
	rs944895	rs944895	rs77172131	rs6062133
	rs76350903	rs76350903	rs944895	rs6142998
	rs2427283	rs2427283	rs76350903	rs77172131
	rs2427284	rs2427284	rs2427283	rs944895
	rs875379	rs875379	rs2427284	rs76350903
	rs114642987	rs114642987	rs875379	rs2427283
	rs13042941	rs13042941	rs114642987	rs2427284
	rs6062223	rs6062223	rs13042941	rs875379
	rs138657380	rs138657380	rs6062223	rs114642987
	rs4925386	rs4925386	rs138657380	rs13042941
	rs4925229	rs4925229	rs4925386	rs6062223
	rs78026347	rs78026347	rs4925229	rs138657380
	rs115914846	rs115914846	rs78026347	rs4925386
	rs3810553	rs3810553	rs115914846	rs4925229
	rs78287067	rs78287067	rs3810553	rs78026347
	rs114526073	rs114526073	rs78287067	rs115914846
	rs111509987	rs111509987	rs114526073	rs3810553
	rs6062251	rs6062251	rs111509987	rs78287067
	rs4635599	rs4635599	rs6062251	rs114526073
			rs4635599	

Appendix Table 5d. SNPs Mapped to Genes within the Chromosome 20 Region of Interest *cont'd.*

Chromosome 20 Genes	***OSBPL2***	***PSMA7***	***SS18L1***	***TAF4***	***CDH4***
Mapped SNPs	rs78979746	rs2024714	rs944260	rs4468878	rs237655
	rs6142884	rs944260	rs78979746	rs4925189	rs2426983
	rs2427254	rs78979746	rs6142884	rs2427158	rs3752252
	rs36106901	rs6142884	rs2427254	rs2024714	rs4468878
	rs6062133	rs2427254	rs36106901	rs944260	rs4925189
	rs6142998	rs36106901	rs6062133	rs78979746	rs2427158
	rs77172131	rs6062133	rs6142998	rs6142884	rs2024714
	rs944895	rs6142998	rs77172131	rs2427254	rs944260
	rs76350903	rs77172131	rs944895	rs36106901	rs78979746
	rs2427283	rs944895	rs76350903	rs6062133	rs6142884
	rs2427284	rs76350903	rs2427283	rs6142998	rs2427254
	rs875379	rs2427283	rs2427284	rs77172131	rs36106901
	rs114642987	rs2427284	rs875379	rs944895	rs6062133
	rs13042941	rs875379	rs114642987	rs76350903	rs6142998
	rs6062223	rs114642987	rs13042941	rs2427283	rs77172131
	rs138657380	rs13042941	rs6062223	rs2427284	rs944895
	rs4925386	rs6062223	rs138657380	rs875379	rs76350903
	rs4925229	rs138657380	rs4925386	rs114642987	rs2427283
	rs78026347	rs4925386	rs4925229	rs13042941	rs2427284
	rs115914846	rs4925229	rs78026347	rs6062223	rs875379
	rs3810553	rs78026347	rs115914846	rs138657380	rs114642987
	rs78287067	rs115914846	rs3810553	rs4925386	rs13042941
	rs114526073	rs3810553	rs78287067	rs4925229	rs6062223
	rs111509987	rs78287067	rs114526073	rs78026347	rs138657380
	rs6062251	rs114526073	rs111509987	rs115914846	rs4925386
	rs4635599	rs111509987	rs6062251	rs3810553	rs4925229
		rs6062251	rs4635599	rs78287067	rs78026347
		rs4635599		rs114526073	rs115914846
					rs3810553
					rs78287067
					rs114526073

Appendix Figure 1a. Triangle Plot of STRUCTURE analysis results for GHAT study participants



Note: Only founder individuals from the HapMap populations shown above were included in STRUCTURE analyses. The following populations are defined above as follows:
GHAT=Current study cohort
CEU = CEU HapMap individuals
YRI = YRI HapMap individuals
CHB + JPT = Combination of all CHB and JPT HapMap individuals

References

1. Global status report on non-communicable diseases 2010. 2011; www.who.int, 2014.

2. Aso Y. Plasminogen activator inhibitor (PAI)-1 in vascular inflammation and thrombosis. *Frontiers in bioscience : a journal and virtual library.* 2007;12:2957-2966.

3. Kohler HP, Grant PJ. Mechanisms of disease: Plasminogen-activator inhibitor type 1 and coronary artery disease. *New Engl J Med.* Jun 15 2000;342(24):1792-1801.

4. Alessi MC, Juhan-Vague I. Contribution of PAI-1 in cardiovascular pathology. *Archives des maladies du coeur et des vaisseaux.* Jun 2004;97(6):673-678.

5. Iwai N, Shimoike H, Nakamura Y, Tamaki S, Kinoshita M. The 4G/5G polymorphism of the plasminogen activator inhibitor gene is associated with the time course of progression to acute coronary syndromes. *Atherosclerosis.* Jan 1998;136(1):109-114.

6. Hong Y, Pedersen NL, Egberg N, de Faire U. Moderate genetic influences on plasma levels of plasminogen activator inhibitor-1 and evidence of genetic and environmental influences shared by plasminogen activator inhibitor-1, triglycerides, and body mass index. *Arterioscler Thromb Vasc Biol.* Nov 1997;17(11):2776-2782.

7. Cesari M, Sartori MT, Patrassi GM, Vettore S, Rossi GP. Determinants of plasma levels of plasminogen activator inhibitor-1 : A study of normotensive twins. *Arterioscler Thromb Vasc Biol.* Feb 1999;19(2):316-320.

8. Kohler HP, Grant PJ. Plasminogen-activator inhibitor type 1 and coronary artery disease. *N Engl J Med.* Jun 15 2000;342(24):1792-1801.

9. Aznar J, Estelles A, Tormo G, et al. Plasminogen-Activator Inhibitor Activity and Other Fibrinolytic Variables in Patients with Coronary-Artery Disease. *Brit Heart J.* May 1988;59(5):535-541.

10. Gram J, Bladbjerg EM, Moller L, Sjol A, Jespersen J. Tissue-type plasminogen activator and C-reactive protein in acute coronary heart disease. A nested case-control study. *Journal of internal medicine.* Feb 2000;247(2):205-212.

11. Pradhan AD, LaCroix AZ, Langer RD, et al. Tissue plasminogen activator antigen and D-dimer as markers for atherothrombotic risk among healthy postmenopausal women. *Circulation.* Jul 20 2004;110(3):292-300.

12. Gorog DA. Prognostic Value of Plasma Fibrinolysis Activation Markers in Cardiovascular Disease. *J Am Coll Cardiol.* Jun 15 2010;55(24):2701-2709.

13. Ridker PM, Brown NJ, Vaughan DE, Harrison DG, Mehta JL. Established and emerging plasma biomarkers in the prediction of first atherothrombotic events. *Circulation.* Jun 29 2004;109(25 Suppl 1):IV6-19.

14. Smith A, Patterson C, Yarnell J, Rumley A, Ben-Shlomo Y, Lowe G. Which hemostatic markers add to the predictive value of conventional risk factors for coronary heart disease and ischemic stroke? The Caerphilly Study. *Circulation.* Nov 15 2005;112(20):3080-3087.

15. Nordt TK, Peter K, Ruef J, Kubler W, Bode C. Plasminogen activator inhibitor type-1 (PAI-1) and its role in cardiovascular disease. *Thromb Haemost.* Sep 1999;82 Suppl 1:14-18.

16. Marcucci R, Brogi D, Sofi F, et al. PAI-1 and homocysteine, but not lipoprotein (a) and thrombophilic polymorphisms, are independently associated with the occurrence of major adverse cardiac events after successful coronary stenting. *Heart.* Mar 2006;92(3):377-381.

17. Schneiderman J, Sawdey MS, Keeton MR, et al. Increased Type-1 Plasminogen-Activator Inhibitor Gene-Expression in Atherosclerotic Human Arteries. *P Natl Acad Sci USA.* Aug 1 1992;89(15):6998-7002.

18. Peetz D, Victor A, Adams P, et al. Genetic and environmental influences on the fibrinolytic system: a twin study. *Thromb Haemost.* Aug 2004;92(2):344-351.

19. Hong Y, Pedersen NL, Egberg N, de Faire U. Moderate genetic influences on plasma levels of plasminogen activator inhibitor-1 and evidence of genetic and environmental influences shared by plasminogen activator inhibitor-1, triglycerides, and body mass index. *Arterioscler Thromb Vasc Biol.* Nov 1997;17(11):2776-2782.

20. Cesari M, Sartori MT, Patrassi GM, Vettore S, Rossi GP. Determinants of plasma levels of plasminogen activator inhibitor-1 : A study of normotensive twins. *Arterioscler Thromb Vasc Biol.* Feb 1999;19(2):316-320.

21. Verschuur M, Jellema A, Bladbjerg EM, et al. The plasminogen activator inhibitor-1 (PAI-1) promoter haplotype is related to PAI-1 plasma concentrations in lean individuals. *Atherosclerosis.* Aug 2005;181(2):275-284.

22. Asselbergs FW, Williams SM, Hebert PR, et al. The gender-specific role of polymorphisms from the fibrinolytic, renin-angiotensin, and bradykinin systems in determining plasma t-PA and PAI-1 levels. *Thromb Haemost.* Oct 2006;96(4):471-477.

23. van der Bom JG, Bots ML, Haverkate F, Kluft C, Grobbee DE. The 4G5G polymorphism in the gene for PAI-1 and the circadian oscillation of plasma PAI-1. *Blood.* Mar 1 2003;101(5):1841-1844.

24. Tsantes AE, Nikolopoulos GK, Bagos PG, et al. Association between the plasminogen activator inhibitor-1 4G/5G polymorphism and venous thrombosis. A meta-analysis. *Thromb Haemost.* Jun 2007;97(6):907-913.

25. Meigs JB, Dupuis J, Liu C, et al. PAI-1 Gene 4G/5G polymorphism and risk of type 2 diabetes in a population-based sample. *Obesity (Silver Spring).* May 2006;14(5):753-758.

26. Yende S, Angus DC, Ding J, et al. 4G/5G plasminogen activator inhibitor-1 polymorphisms and haplotypes are associated with pneumonia. *Am J Respir Crit Care Med.* Dec 1 2007;176(11):1129-1137.

27. Kathiresan S, Gabriel SB, Yang Q, et al. Comprehensive survey of common genetic variation at the plasminogen activator inhibitor-1 locus and relations to circulating plasminogen activator inhibitor-1 levels. *Circulation.* Sep 20 2005;112(12):1728-1735.

28. Penrod NM, Poku KA, Vaughan DE, et al. Epistatic interactions in genetic regulation of t-PA and PAI-1 levels in a Ghanaian population. *PLoS One.* 2011;6(1):e16639.

29. Asselbergs FW, Williams SM, Hebert PR, et al. Epistatic effects of polymorphisms in genes from the renin-angiotensin, bradykinin, and fibrinolytic systems on plasma t-PA and PAI-1 levels. *Genomics.* Mar 2007;89(3):362-369.

30. Asselbergs FW, Williams SM, Hebert PR, et al. The effects of polymorphisms in genes from the renin-angiotensin, bradykinin, and fibrinolytic systems on plasma t-PA and PAI-1 levels are dependent on environmental context. *Hum Genet.* Nov 2007;122(3-4):275-281.

31. Huang J, Sabater-Lleal M, Asselbergs FW, et al. Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. *Blood.* Dec 6 2012;120(24):4873-4881.

32. Mukamal KJ, Jadhav PP, D'Agostino RB, et al. Alcohol consumption and hemostatic factors: analysis of the Framingham Offspring cohort. *Circulation.* Sep 18 2001;104(12):1367-1373.

33. Appel SJ, Harrell JS, Davenport ML. Central obesity, the metabolic syndrome, and plasminogen activator inhibitor-1 in young adults. *J Am Acad Nurse Pract.* Dec 2005;17(12):535-541.

34. Kenny S, Gamble J, Lyons S, et al. Gastric expression of plasminogen activator inhibitor (PAI)-1 is associated with hyperphagia and obesity in mice. *Endocrinology.* Feb 2013;154(2):718-726.

35. Berberoglu M, Evliyaoglu O, Adiyaman P, et al. Plasminogen activator inhibitor-1 (PAI-1) gene polymorphism (-675 4G/5G) associated with obesity and vascular risk in children. *J Pediatr Endocrinol Metab.* May 2006;19(5):741-748.

36. Djousse L, Pankow JS, Arnett DK, et al. Alcohol consumption and plasminogen activator inhibitor type 1: the National Heart, Lung, and Blood Institute Family Heart Study. *Am Heart J.* Apr 2000;139(4):704-709.

37. Taivainen H, Laitinen K, Tahtela R, Kilanmaa K, Valimaki MJ. Role of plasma vasopressin in changes of water balance accompanying acute alcohol intoxication. *Alcoholism, clinical and experimental research.* Jun 1995;19(3):759-762.

38. Thevananther S, Brecher AS. Interaction of acetaldehyde with plasma proteins of the renin-angiotensin system. *Alcohol.* Nov-Dec 1994;11(6):493-499.

39. Skurk T, Hauner H. Obesity and impaired fibrinolysis: role of adipose production of plasminogen activator inhibitor-1. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity.* Nov 2004;28(11):1357-1364.

40. Sartori MT, Vettor R, De Pergola G, et al. Role of the 4G/5G polymorphism of PaI-1 gene promoter on PaI-1 levels in obese patients: influence of fat distribution and insulin-resistance. *Thromb Haemost.* Nov 2001;86(5):1161-1169.

41. Ma LJ, Mao SL, Taylor KL, et al. Prevention of obesity and insulin resistance in mice lacking plasminogen activator inhibitor 1. *Diabetes.* Feb 2004;53(2):336-346.

42. Pieters M, de Lange Z, Hoekstra T, Ellis SM, Kruger A. Triglyceride concentration and waist circumference influence alcohol-related plasminogen activator inhibitor-1 activity increase in black South Africans. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis.* Dec 2010;21(8):736-743.

43. Bastard JP, Pieroni L. Plasma plasminogen activator inhibitor 1, insulin resistance and android obesity. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie.* Dec 1999;53(10):455-461.

44. Bastard JP, Pieroni L, Hainque B. Relationship between plasma plasminogen activator inhibitor 1 and insulin resistance. *Diabetes/metabolism research and reviews.* May-Jun 2000;16(3):192-201.

45. Juhan-Vague I, Alessi MC. PAI-1, obesity, insulin resistance and risk of cardiovascular events. *Thrombosis and haemostasis.* Jul 1997;78(1):656-660.

46. Juhan-Vague I, Alessi MC, Morange PE. Hypofibrinolysis and increased PAI-1 are linked to atherothrombosis via insulin resistance and obesity. *Ann Med.* Dec 2000;32 Suppl 1:78-84.

47. Brown NJ, Murphey LJ, Srikuma N, Koschachuhanan N, Williams GH, Vaughan DE. Interactive effect of PAI-1 4G/5G genotype and salt intake on PAI-1 antigen. *Arterioscler Thromb Vasc Biol.* Jun 2001;21(6):1071-1077.

48. Williams SM, Stocki S, Jiang L, et al. A population-based study in Ghana to investigate inter-individual variation in plasma t-PA and PAI-1. *Ethn Dis.* Summer 2007;17(3):492-497.

49. Byberg L, Siegbahn A, Berglund L, McKeigue P, Reneland R, Lithell H. Plasminogen activator inhibitor-1 activity is independently related to both insulin sensitivity and serum triglycerides in 70-year-old men. *Arterioscler Thromb Vasc Biol.* Feb 1998;18(2):258-264.

50. Alessi MC, Juhan-Vague I. [Endothelium, thrombosis and fibrinolysis]. *La Revue du praticien.* Dec 15 1997;47(20):2227-2231.

51. Cesari M, Pahor M, Incalzi RA. Plasminogen activator inhibitor-1 (PAI-1): a key factor linking fibrinolysis and age-related subclinical and clinical conditions. *Cardiovascular therapeutics.* Oct 2010;28(5):e72-91.

52. Sobel BE, Lee YH, Pratley RE, Schneider DJ. Increased plasminogen activator inhibitor type-1 (PAI-1) in the heart as a function of age. *Life sciences.* Sep 20 2006;79(17):1600-1605.

53. Ardite E, Perdiguero E, Vidal B, Gutarra S, Serrano AL, Munoz-Canoves P. PAI-1-regulated miR-21 defines a novel age-associated fibrogenic pathway in muscular dystrophy. *J Cell Biol.* Jan 9 2012;196(1):163-175.

54. Harslund J, Nielsen OL, Brunner N, Offenberg H. Gender-dependent physiological implications of combined PAI-1 and TIMP-1 gene deficiency characterized in a mouse model. *Am J Physiol Regul Integr Comp Physiol.* Oct 2007;293(4):R1630-1639.

55. Asselbergs FW, Williams SM, Hebert PR, et al. The gender-specific role of polymorphisms from the fibrinolytic, renin-angiotensin, and bradykinin systems in determining plasma t-PA and PAI-1 levels. *Thrombosis and haemostasis.* Oct 2006;96(4):471-477.

56. Asselbergs FW, Williams SM, Hebert PR, et al. Gender-specific correlations of plasminogen activator inhibitor-1 and tissue plasminogen activator levels with cardiovascular disease-related traits. *J Thromb Haemost.* Feb 2007;5(2):313-320.

57. Schoenhard JA, Asselbergs FW, Poku KA, et al. Male-female differences in the genetic regulation of t-PA and PAI-1 levels in a Ghanaian population. *Hum Genet.* Dec 2008;124(5):479-488.

58. Williams SM, Stocki S, Jiang L, et al. A population-based study in Ghana to investigate inter-individual variation in plasma t-PA and PAI-1. *Ethn Dis.* Summer 2007;17(3):492-497.

59. Schoenhard JA, Asselbergs FW, Poku KA, et al. Male-female differences in the genetic regulation of t-PA and PAI-1 levels in a Ghanaian population. *Hum Genet.* Dec 2008;124(5):479-488.

60. Rona G. The pathogenesis of human myocardial infarction. *Canadian Medical Association journal.* Nov 12 1966;95(20):1012-1019.

61. Samad F, Ruf W. Inflammation, obesity, and thrombosis. *Blood.* Nov 14 2013;122(20):3415-3422.

62. Booth NA, Bennett B. Fibrinolysis and thrombosis. *Bailliere's clinical haematology.* Sep 1994;7(3):559-572.

63. Kawasaki T, Dewerchin M, Lijnen HR, Vermylen J, Hoylaerts MF. Vascular release of plasminogen activator inhibitor-1 impairs fibrinolysis during acute arterial thrombosis in mice. *Blood.* Jul 1 2000;96(1):153-160.

64. Saidi S, Slamia LB, Mahjoub T, Ammou SB, Almawi WY. Association of PAI-1 4G/5G and -844G/A gene polymorphism and changes in PAI-1/tPA levels in stroke: a case-control study. *Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association.* Jul-Aug 2007;16(4):153-159.

65. Abboud N, Ghazouani L, Saidi S, et al. Association of PAI-1 4G/5G and -844G/A gene polymorphisms and changes in PAI-1/tissue plasminogen activator levels in myocardial infarction: a case-control study. *Genetic testing and molecular biomarkers.* Feb 2010;14(1):23-27.

66. Koch W, Schrempf M, Erl A, et al. 4G/5G polymorphism and haplotypes of SERPINE1 in atherosclerotic diseases of coronary arteries. *Thrombosis and haemostasis.* Jun 2010;103(6):1170-1180.

67. Onalan O, Balta G, Oto A, et al. Plasminogen activator inhibitor-1 4G4G genotype is associated with myocardial infarction but not with stable coronary artery disease. *Journal of thrombosis and thrombolysis.* Dec 2008;26(3):211-217.

68. Juhan-Vague I, Alessi MC, Mavri A, Morange PE. Plasminogen activator inhibitor-1, inflammation, obesity, insulin resistance and vascular risk. *J Thromb Haemost.* Jul 2003;1(7):1575-1579.

69. Juhan-Vague I, Morange PE, Frere C, et al. The plasminogen activator inhibitor-1 -675 4G/5G genotype influences the risk of myocardial infarction associated with elevated plasma proinsulin and insulin concentrations in men from Europe: the HIFMECH study. *J Thromb Haemost.* Nov 2003;1(11):2322-2329.

70. Zhan M, Zhou Y, Han Z. Plasminogen activator inhibitor-1 4G/5G gene polymorphism in patients with myocardial or cerebrovascular infarction in Tianjin, China. *Chin Med J (Engl).* Nov 2003;116(11):1707-1710.

71. Grant PJ. Plasminogen activator inhibitor-1 gene and myocardial infarction. *Circulation.* Nov 18 1997;96(10):3796-3797.

72. Ridker PM, Hennekens CH, Lindpaintner K, Stampfer MJ, Miletich JP. Arterial and venous thrombosis is not associated with the 4G/5G polymorphism in the promoter of the plasminogen activator inhibitor gene in a large cohort of US men. *Circulation.* Jan 7 1997;95(1):59-62.

73. Tsantes AE, Nikolopoulos GK, Bagos PG, et al. Association between the plasminogen activator inhibitor-1 4G/5G polymorphism and venous thrombosis. A meta-analysis. *Thrombosis and haemostasis.* Jun 2007;97(6):907-913.

74. Tsantes AE, Nikolopoulos GK, Bagos PG, et al. Plasminogen activator inhibitor-1 4G/5G polymorphism and risk of ischemic stroke: a meta-analysis. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis.* Jul 2007;18(5):497-504.

75. Attia J, Thakkinstian A, Wang Y, et al. The PAI-1 4G/5G gene polymorphism and ischemic stroke: an association study and meta-analysis. *Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association.* Jul-Aug 2007;16(4):173-179.

76. Boekholdt SM, Bijsterveld NR, Moons AH, Levi M, Buller HR, Peters RJ. Genetic variation in coagulation and fibrinolytic proteins and their relation with acute myocardial infarction: a systematic review. *Circulation.* Dec 18 2001;104(25):3063-3068.

77. de Lange M, Snieder H, Ariens RA, Spector TD, Grant PJ. The genetics of haemostasis: a twin study. *Lancet.* Jan 13 2001;357(9250):101-105.

78. Peetz D, Victor A, Adams P, et al. Genetic and environmental influences on the fibrinolytic system: a twin study. *Thrombosis and haemostasis.* Aug 2004;92(2):344-351.

79. Asselbergs FW, Pattin K, Snieder H, Hillege HL, van Gilst WH, Moore JH. Genetic architecture of tissue-type plasminogen activator and plasminogen activator inhibitor-1. *Seminars in thrombosis and hemostasis.* Sep 2008;34(6):562-568.

80. Henry M, Tregouet DA, Alessi MC, et al. Metabolic determinants are much more important than genetic polymorphisms in determining the PAI-1 activity and antigen plasma concentrations: a family study with part of the Stanislas Cohort. *Arterioscler Thromb Vasc Biol.* Jan 1998;18(1):84-91.

81. Schoenhard JA, Smith LH, Painter CA, Eren M, Johnson CH, Vaughan DE. Regulation of the PAI-1 promoter by circadian clock components: differential activation by BMAL1 and BMAL2. *J Mol Cell Cardiol.* May 2003;35(5):473-481.

82. Gong LL, Peng JH, Han FF, et al. Association of tissue plasminogen activator and plasminogen activator inhibitor polymorphism with myocardial infarction: a meta-analysis. *Thromb Res.* Sep 2012;130(3):e43-51.

83. Nikolopoulos GK, Bagos PG, Tsangaris I, et al. The association between plasminogen activator inhibitor type 1 (PAI-1) levels, PAI-1 4G/5G polymorphism, and myocardial infarction: a Mendelian randomization meta-analysis. *Clin Chem Lab Med.* Jul 2014;52(7):937-950.

84. Koenker RB, G. Regression quantiles. *Econometrica*1978:33-50.

85. Briollais L, Durrieu G. Application of quantile regression to recent genetic and -omic studies. *Hum Genet.* Apr 26 2014.

86. Koenker R. *Quantile Regression*. New York: Cambridge University Press; 2005.

87. John ON, E. Quantile Regression Analysis as a robust alternative to ordinary least squares. *Scientia Africana.* 2009;8(2):61-65.

88. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics.* Sep 2007;81(3):559-575.

89. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* Jun 2000;155(2):945-959.

90. *Stata Statistical Software: Release 11.* [computer program]. College Station, TX: StataCorp LP; 2009.

91. *R: A Language and Environment for Statistical Computing* [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2013.

92. Braun-Fahrlander C, Wuthrich B, Gassner M, et al. Validation of a rhinitis symptom questionnaire (ISAAC core questions) in a population of Swiss school children visiting the school health services. SCARPOL-team. Swiss Study on Childhood Allergy and Respiratory Symptom with respect to Air Pollution and Climate. International Study of Asthma and Allergies in Childhood. *Pediatr Allergy Immunol.* May 1997;8(2):75-82.

93. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* Sep 15 2010;26(18):2336-2337.

94. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological).* 1995;57:289-300.

95. Xu Z, Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* Jul 2009;37(Web Server issue):W600-605.

96. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet.* Mar 15 2001;10(6):591-597.

97. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics.* 2006;7:166.

98. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* Jul 1 2003;31(13):3576-3579.

99. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* Jul 1 2003;31(13):3568-3571.

100. Bhattacharyya S, Tobacman JK. Molecular signature of kappa-carrageenan mimics chondroitin-4-sulfate and dermatan sulfate and enables interaction with arylsulfatase B. *The Journal of nutritional biochemistry.* Sep 2012;23(9):1058-1063.

101. Bhattacharyya S, Tobacman JK. Hypoxia reduces arylsulfatase B activity and silencing arylsulfatase B replicates and mediates the effects of hypoxia. *PloS one.* 2012;7(3):e33250.

102. Sorensen BS, Toustrup K, Horsman MR, Overgaard J, Alsner J. Identifying pH independent hypoxia induced genes in human squamous cell carcinomas in vitro. *Acta oncologica.* Oct 2010;49(7):895-905.

103. Normant E, Gros C, Schwartz JC. Carboxypeptidase A isoforms produced by distinct genes or alternative splicing in brain and other extrapancreatic tissues. *J Biol Chem.* Sep 1 1995;270(35):20543-20549.

104. Bentley L, Nakabayashi K, Monk D, et al. The imprinted region on human chromosome 7q32 extends to the carboxypeptidase A gene cluster: an imprinted candidate for Silver-Russell syndrome. *Journal of medical genetics.* Apr 2003;40(4):249-256.

105. Pereira HJ, Souza LL, Costa-Neto CM, Salgado MC, Oliveira EB. Carboxypeptidases A1 and A2 from the perfusate of rat mesenteric arterial bed differentially process angiotensin peptides. *Peptides.* Jan 2012;33(1):67-76.

106. Ocaranza MP, Jalil JE. Protective Role of the ACE2/Ang-(1-9) Axis in Cardiovascular Remodeling. *International journal of hypertension.* 2012;2012:594361.

107. Skurk T, Lee YM, Rohrig K, Hauner H. Effect of angiotensin peptides on PAI-1 expression and production in human adipocytes. *Horm Metab Res.* Apr 2001;33(4):196-200.

108. Skurk T, Lee YM, Hauner H. Angiotensin II and its metabolites stimulate PAI-1 protein release from human adipocytes in primary culture. *Hypertension.* May 2001;37(5):1336-1340.

109. Vaughan DE, Lazos SA, Tong K. Angiotensin II regulates the expression of plasminogen activator inhibitor-1 in cultured endothelial cells. A potential link between the renin-angiotensin system and thrombosis. *J Clin Invest.* Mar 1995;95(3):995-1001.

110. Brown NJ, Agirbasli M, Vaughan DE. Comparative effect of angiotensin-converting enzyme inhibition and angiotensin II type 1 receptor antagonism on plasma fibrinolytic balance in humans. *Hypertension.* Aug 1999;34(2):285-290.

111. Brown NJ, Kim KS, Chen YQ, et al. Synergistic effect of adrenal steroids and angiotensin II on plasminogen activator inhibitor-1 production. *The Journal of clinical endocrinology and metabolism.* Jan 2000;85(1):336-344.

112. Brown NJ, Vaughan DE. Prothrombotic effects of angiotensin. *Advances in internal medicine.* 2000;45:419-429.

113. Wende H, Volz A, Ziegler A. Extensive gene duplications and a large inversion characterize the human leukocyte receptor cluster. *Immunogenetics.* Jul 2000;51(8-9):703-713.

114. Williams SM, Haines JL. Correcting away the hidden heritability. *Ann Hum Genet.* May 2011;75(3):348-350.

115. Mitchell JA, Hakonarson H, Rebbeck TR, Grant SF. Obesity-susceptibility loci and the tails of the pediatric BMI distribution. *Obesity.* Jun 2013;21(6):1256-1260.

116. Williams PT. Quantile-specific penetrance of genes affecting lipoproteins, adiposity and height. *PloS one.* 2012;7(1):e28764.

117. Huang L, Zhu W, Saunders CP, et al. A novel application of quantile regression for identification of biomarkers exemplified by equine cartilage microarray data. *BMC Bioinformatics.* 2008;9:300.

118. Beyerlein A, von Kries R, Ness AR, Ong KK. Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children. *PloS one.* 2011;6(4):e19057.

119. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* Jan 15 2005;21(2):263-265.

120. Philippe C, Porter DE, Emerton ME, Wells DE, Simpson AH, Monaco AP. Mutation screening of the EXT1 and EXT2 genes in patients with hereditary multiple exostoses. *American journal of human genetics.* Sep 1997;61(3):520-528.

121. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature.* Feb 22 2007;445(7130):881-885.

122. Liu L, Yang X, Wang H, et al. Association between variants of EXT2 and type 2 diabetes: a replication and meta-analysis. *Hum Genet.* Feb 2013;132(2):139-145.

123. Konkle BA, Kollros PR, Kelly MD. Heparin-binding growth factor-1 modulation of plasminogen activator inhibitor-1 expression. Interaction with cAMP and protein kinase C-mediated pathways. *J Biol Chem.* Dec 15 1990;265(35):21867-21873.

124. Guess J, Burch JB, Ogoussan K, et al. Circadian disruption, Per3, and human cytokine secretion. *Integrative cancer therapies.* Dec 2009;8(4):329-336.

125. Cheng B, Anea CB, Yao L, et al. Tissue-intrinsic dysfunction of circadian clock confers transplant arteriosclerosis. *Proc Natl Acad Sci U S A.* Oct 11 2011;108(41):17147-17152.

126. Ko CH, Takahashi JS. Molecular components of the mammalian circadian clock. *Hum Mol Genet.* Oct 15 2006;15 Spec No 2:R271-277.

127. Chong NW, Codd V, Chan D, Samani NJ. Circadian clock genes cause activation of the human PAI-1 gene promoter with 4G/5G allelic preference. *FEBS Lett.* Aug 7 2006;580(18):4469-4472.

128. Stow LR, Gumz ML. The circadian clock in the kidney. *J Am Soc Nephrol.* Apr 2011;22(4):598-604.

129. Ciarleglio CM, Ryckman KK, Servick SV, et al. Genetic differences in human circadian clock genes among worldwide populations. *Journal of biological rhythms.* Aug 2008;23(4):330-340.

130. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & development.* Mar 1 2006;20(5):515-524.

131. Baumann FC, Boizard-Callais F, Labat-Robert J. Trehalase activity in genetically diabetic mice (serum, kidney, and liver). *Journal of medical genetics.* Dec 1981;18(6):418-423.

132. Isichei UP, Gorecki T. Serum trehalase activities in controlled and uncontrolled diabetes and the impact of oral glucose, high carbohydrate and glycosuria on serum levels. *African journal of medicine and medical sciences.* Jun 1993;22(2):5-11.

133. Muller YL, Hanson RL, Knowler WC, et al. Identification of genetic variation that determines human trehalase activity and its association with type 2 diabetes. *Hum Genet.* Jun 2013;132(6):697-707.

134. Ouyang Y, Xu Q, Mitsui K, Motizuki M, Xu Z. Human trehalase is a stress responsive protein in Saccharomyces cerevisiae. *Biochem Biophys Res Commun.* Feb 6 2009;379(2):621-625.

135. Iwadate Y, Hayama M, Adachi A, et al. High serum level of plasminogen activator inhibitor-1 predicts histological grade of intracerebral gliomas. *Anticancer research.* Jan-Feb 2008;28(1B):415-418.

136. Gao X, Mi Y, Yan A, et al. The PHLDB1 rs498872 (11q23.3) polymorphism and glioma risk: A meta-analysis. *Asia-Pacific journal of clinical oncology.* Jun 17 2014.

137. Vaughan DE. PAI-1 and atherothrombosis. *J Thromb Haemost.* Aug 2005;3(8):1879-1883.

138. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet.* Jun 2009;10(6):392-404.

139. Gilbert-Diamond D, Moore JH. Analysis of gene-gene interactions. *Curr Protoc Hum Genet.* Jul 2011;Chapter 1:Unit1 14.

140. Aad G, Abajyan T, Abbott B, et al. Observation of associated near-side and away-side long-range correlations in sqrt[s(NN)]=5.02 TeV proton-lead collisions with the ATLAS detector. *Physical review letters.* May 3 2013;110(18):182302.

141. Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* Oct 2000;10(10):1435-1444.

142. Chen Y, Rollins J, Paigen B, Wang X. Genetic and genomic insights into the molecular basis of atherosclerosis. *Cell metabolism.* Sep 2007;6(3):164-179.

143. Gui J, Moore JH, Williams SM, et al. A Simple and Computationally Efficient Approach to Multifactor Dimensionality Reduction Analysis of Gene-Gene Interactions for Quantitative Traits. *PloS one.* 2013;8(6):e66545.

144. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res.* Jun 2002;12(6):996-1006.

145. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American journal of human genetics.* Jul 2001;69(1):138-147.

146. Imperatore G, Knowler WC, Pettitt DJ, et al. A locus influencing total serum cholesterol on chromosome 19p: results from an autosomal genomic scan of serum lipid concentrations in Pima Indians. *Arterioscler Thromb Vasc Biol.* Dec 2000;20(12):2651-2656.

147. Feitosa MF, Borecki IB, Rich SS, et al. Quantitative-trait loci influencing body-mass index reside on chromosomes 7 and 13: the National Heart, Lung, and Blood Institute Family Heart Study. *American journal of human genetics.* Jan 2002;70(1):72-82.

148. Wang Q, Rao S, Shen GQ, et al. Premature myocardial infarction novel susceptibility locus on chromosome 1P34-36 identified by genomewide linkage analysis. *American journal of human genetics.* Feb 2004;74(2):262-271.

149. Wilson SG, Adam G, Langdown M, et al. Linkage and potential association of obesity-related phenotypes with two genes on chromosome 12q24 in a female dizygous twin cohort. *European journal of human genetics : EJHG.* Mar 2006;14(3):340-348.

150. Arya R, Duggirala R, Almasy L, et al. Linkage of high-density lipoprotein-cholesterol concentrations to a locus on chromosome 9p in Mexican Americans. *Nature genetics.* Jan 2002;30(1):102-105.

151. Tejero ME, Cai G, Goring HH, et al. Linkage analysis of circulating levels of adiponectin in Hispanic children. *Int J Obes (Lond).* Mar 2007;31(3):535-542.

152. Meyre D, Lecoeur C, Delplanque J, et al. A genome-wide scan for childhood obesity-associated traits in French families shows significant linkage on chromosome 6q22.31-q23.2. *Diabetes.* Mar 2004;53(3):803-811.

153. Norris JM, Langefeld CD, Scherzinger AL, et al. Quantitative trait loci for abdominal fat and BMI in Hispanic-Americans and African-Americans: the IRAS Family study. *Int J Obes (Lond).* Jan 2005;29(1):67-77.

154. Wiltshire S, Hattersley AT, Hitman GA, et al. A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *American journal of human genetics.* Sep 2001;69(3):553-569.

155. Li WD, Dong C, Li D, Garrigan C, Price RA. A genome scan for serum triglyceride in obese nuclear families. *J Lipid Res.* Mar 2005;46(3):432-438.

156. Lee JH, Reed DR, Li WD, et al. Genome scan for human obesity and linkage to markers in 20q13. *American journal of human genetics.* Jan 1999;64(1):196-209.

157. Luo TH, Zhao Y, Li G, et al. A genome-wide search for type II diabetes susceptibility genes in Chinese Hans. *Diabetologia.* Apr 2001;44(4):501-506.

158. Laulederkind SJ, Hayman GT, Wang SJ, et al. The Rat Genome Database 2013--data, tools and users. *Briefings in bioinformatics.* Jul 2013;14(4):520-526.

159. van Leeuwen C, Tops C, Breukel C, van der Klift H, Fodde R, Khan PM. CA repeat polymorphism at the D5S299 locus linked to adenomatous polyposis coli (APC). *Nucleic Acids Res.* Oct 25 1991;19(20):5805.

160. Bapat B, Mitri A, Greenberg CR. Improved predictive carrier testing for familial adenomatous polyposis using DNA from a single archival specimen and polymorphic markers with multiple alleles. *Hum Pathol.* Dec 1993;24(12):1376-1379.

161. Horii A, Nakatsuru S, Ichii S, Nagase H, Nakamura Y. Multiple forms of the APC gene transcripts and their tissue-specific expression. *Hum Mol Genet.* Mar 1993;2(3):283-287.

162. Mutoh M, Niho N, Komiya M, et al. Plasminogen activator inhibitor-1 (Pai-1) blockers suppress intestinal polyp formation in Min mice. *Carcinogenesis.* Apr 2008;29(4):824-829.

163. Dimova EY, Jakubowska MM, Kietzmann T. CREB binding to the hypoxia-inducible factor-1 responsive elements in the plasminogen activator inhibitor-1 promoter mediates the glucagon effect. *Thrombosis and haemostasis.* Aug 2007;98(2):296-303.

164. Dimova EY, Samoylenko A, Kietzmann T. FOXO4 induces human plasminogen activator inhibitor-1 gene expression via an indirect mechanism by modulating HIF-1alpha and CREB levels. *Antioxidants & redox signaling.* Aug 15 2010;13(4):413-424.

165. Larabee JL, Shakir SM, Hightower L, Ballard JD. Adenomatous polyposis coli protein associates with C/EBP beta and increases Bacillus anthracis edema toxin-stimulated gene expression in macrophages. *J Biol Chem.* Jun 3 2011;286(22):19364-19372.

166. He W, Tan R, Dai C, et al. Plasminogen activator inhibitor-1 is a transcriptional target of the canonical pathway of Wnt/beta-catenin signaling. *J Biol Chem.* Aug 6 2010;285(32):24665-24675.

167. Ishiguro H, Furukawa Y, Daigo Y, et al. Isolation and characterization of human NBL4, a gene involved in the beta-catenin/tcf signaling pathway. *Japanese journal of cancer research : Gann.* Jun 2000;91(6):597-603.

168. Kolligs FT, Bommer G, Goke B. Wnt/beta-catenin/tcf signaling: a critical pathway in gastrointestinal tumorigenesis. *Digestion.* 2002;66(3):131-144.